PB84-115062

Evaluation of Performance Measures for an Urban Photochemical Model

National Center for Atmospheric Research
Boulder, CO

Prepared for

Environmental Protection Agency
Research Triangle Park, NC

Jul 83

U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA 450/4-83-021	2.	3. RECIPIENT'S ACCESSION NO. PB84 115062
4. TITLE AND SUBTITLE Evaluation of Performance Measures for an Urban Photochemical Model		5. REPORT DATE July 1983
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Robin L. Dennis, Mary W. Downton and Robbi S. Keil		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS National Center for Atmospheric Research Environmental and Societal Impacts Group P.O. Box 3000 Boulder, Colorado 80307		10. PROGRAM ELEMENT NO. A13A2A
		11. CONTRACT/GRANT NO. AD-49-F-0-167-0
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency Office of Air Quality Planning and Standards Monitoring and Data Analysis Division (MD-14) Research Triangle Park, North Carolina 27711		13. TYPE OF REPORT AND PERIOD COVERED
		14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

Henry S. Cole, Project Officer

16. ABSTRACT

A workshop conducted by the American Meteorological Society for EPA in September 1980 recommended a large set of statistical measures for use in the evaluation of air quality models. The present study was designed to test the recommended measures in an actual performance evaluation of an airshed model on data developed for Denver, Colorado. Three versions of the SAI Urban Airshed Model were examined. The study involved both an evaluation of the models and an evaluation of the statistical performance measures. The evaluation of the models had two parts--a base year case and an emissions trend case, the latter representing the use of the models for regulatory purposes. Resulting recommendations are intended to aid the future use of such models, the planning of future performance evaluations of airshed models, and the use of performance evaluation statistics.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air pollution Atmospheric models Photochemical reactions Smog Ozone Nitrogen Oxides Hydrocarbons	Urban Airshed Model SAI Airshed Model Carbon-Bond Mechanism Denver	

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 219
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-83-021
July 1983

PB84-115062

Air

Evaluation of Performance Measures for an Urban Photochemical Model

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA-450/4-83-021

Evaluation of Performance Measures for an Urban Photochemical Model

Robin L. Dennis,
Mary W. Downton
and
Robbi S. Keil

by

National Center for Atmospheric Research
Environmental and Societal Impacts Group
P.O. Box 3000
Boulder, Colorado 80307

Contract No. AD-49-F-0-167-0

Prepared for

U.S. ENVIRONMENTAL PROTECTION AGENCY
Office of Air Quality Planning and Standards
Monitoring and Data Analysis Division (MD-14)
Research Triangle Park, NC 27711

July 1983

DISCLAIMER

This report has been reviewed by the Office of Air Quality Planning and Standards, U. S. Environmental Protection Agency, and approved for publication as received from National Center for Atmospheric Research. Approval does not signify that the contents necessarily reflect the views and policies of the U. S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. Copies of this report are available from the National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161.

EVALUATION OF PERFORMANCE MEASURES FOR AN URBAN PHOTOCHEMICAL MODEL

EXECUTIVE SUMMARY

The AMS/EPA Dispersion Model Performance Workshop, in September 1980, recommended a large set of statistical measures for use in the evaluation of air quality models. The present study was designed to test the recommended measures in an actual performance evaluation on Denver data, using three versions of the SAI Urban Airshed Model, termed DOT, EPA1 and EPA$_2$§. The study involved both an evaluation of the models and an evaluation of the statistical performance measures. The evaluation of the models had two parts--a base year case and an emissions trend case. Resulting recommendations are intended to aid in the future use of the models and in the planning of future performance evaluations on urban airshed models.

Evaluation of the Models: Base Year Case

The three models in this study represent successive improvements in the photochemical airshed model. All three versions showed considerable bias (systematic underprediction) and noise, and a variety of errors. We were able to identify several types of errors that degraded the models' performance. They were: missing the peak in space, missing the peak in time, too rapid a vertical mixing, errors introduced by some of the model inputs, and difficulty in treating concentrated sources of NO_x emissions. There continued to be systematic errors that contribute to chronic underprediction that could not be identified. [illegible] seems that this model will have a problem with missing the peak in space and time for typical regulatory cases in which a peak has been observed at a monitoring site, particularly when there are few monitoring stations. The predicted peaks are, however, in approximately the correct locations. The model randomly misses the peak in time, within two-hour limits, but this is judged not to be a significant problem for regulatory use of the model.

The models only became differentiated by their peak predictions. The oldest model version, DOT, was the worst and the newest model version, EPA2, was the best. There was still bias (underprediction) in the peak ozone predictions of EPA2, not less than 10% and not more than 30%. The responsiveness of peak ozone predictions of EPA2 to changes in meteorology appears to be less than is actually observed. The DOT model appeared to respond randomly to changes in meteorology. There were too few monitoring stations to assess the size of the predicted "ozone cloud". Sizeable clouds were predicted, but our impression is that they are smaller than the observed "clouds". Based on some of the systematic errors identified, it appears that the model can still be improved.

Evaluation of the Models: Emissions Trend Case

In the course of this study, it became clear that a performance evaluation using a data set in which the emissions do not change cannot provide reliable inferences about the performance of the model predictions when the emissions do change. Thus if the model is to be used to predict changes in ozone concentrations due to changes in emissions, then the model predictions must be tested with a data set in which a change in emissions and a corresponding change in ambient ozone concentrations has been established.

For the prediction of the trend in peak ozone due to a change in emissions, EPA2 was again the best. It did appear to underpredict the change in peak ozone--13 percent observed versus 10 percent predicted change over a 3-year period. The data base was too small to make any firm estimate, however. The Urban Airshed Model has certain idiosyncrasies that affect its predictions for regulatory use. In particular, its predictions of a change in peak ozone due to a change in emissions is affected by the vertical mixing rate. It appears that EPA2 can still be improved.

Bias estimates from a base year evaluation do not seem to be adequate indicators of how well the model predicts trends in ozone resulting from a change of emissions. Some errors affect both base year predictions and ozone trend predictions, but the predictions for the two cases are not equally sensitive to the same error. Other errors seem to affect only the base year predictions. This suggests that if certain errors can be fixed, then it is possible that the existence of bias in the model's predictions for a historical day will not significantly affect its prediction of a relative change in peak ozone due to a change in emissions. Thus EPA2 seems to be more adequate from a regulatory perspective than from a purely scientific perspective.

Evaluation of the Performance Measures

A performance evaluation should be structured around attributes which are important in the use of the model. Performance measures should then be chosen on the basis of the attributes that have been selected. Thus the list of measures and comparisons recommended by the AMS/EPA workshop included some measures which were not appropriate for the Denver application and failed to include a measure of the response to emissions change which was needed in the Denver application. In addition, many graphical displays not mentioned at the workshop were found to be extremely useful. Measures found most useful in this study were bias, noise, absolute deviation, and correlation-related statistics, applied to subsets of the Denver data. Emphasis was placed on comparisons of completely paired hourly data for a diagnostic understanding of the models and on comparisons of the daily maxima for regulatory purposes.

For detailed analysis of the locations and causes of errors in the models, statistics computed separately for each hour and site, or for each day and site, were most helpful. Sensitivity analyses and graphical analyses of model performance under controlled changes in the data and in the model were necessary to further explain the errors. Analyses of daily peak concentrations revealed additional information because they were sensitive to different aspects of model performance. To evaluate the models' performance for regulatory purposes, statistics computed on the daily maximum predictions were most appropriate.

The performance measures were found to be useful aids in comparing models, but subgrouping of the data, graphical analysis, sensitivity analysis and case-by-case analysis was necessary for diagnosing errors in models. Thus, it was felt that the measures would be inappropriate as absolute performance standards. Understanding of the reasons underlying the computed measures was necessary for a meaningful evaluation, and professional judgment was required in drawing conclusions. We expect this to be typical with air quality models. The evaluation of a model will not be a simple, routine matter. The statistical measures provide an aid by defining consistent "vital statistics" for a model. But they are not a substitute for the detailed, diagnostic analysis necessary to support an evaluation of the adequacy of a model for use in decision-making.

---

§DOT: Urban Airshed Model with Carbon Bond I chemistry; EPA1: Urban Airshed Model with Carbon Bond II chemistry; EPA2: Urban Airshed Model with Carbon Bond II and revised numerical algorithm.

TABLE OF CONTENTS

Page

I. INTRODUCTION 1
  A. SELECTION OF EVALUATION CRITERIA 2
    Evaluation of a Model for Scientific or Regulatory Purposes 3
    Requirements Due to the Type of Model 4
    Requirements Based on Use of the Model 4
    Model Attributes Selected for the Denver Performance Evaluation 6
  B. PRACTICAL LIMITATIONS ON PERFORMANCE EVALUATION 6

II. DISCUSSION OF THE PERFORMANCE MEASURES. 8
  A. BASIC ASSUMPTIONS OF STATISTICAL TESTING 8
    Choice of a Data Sample for Model Evaluation 9
    Normality of the Data 9
    Lack of Independence 11
    Other Sources of Dependence 14

  B. PERFORMANCE MEASURES 15
    Gross Error 16
    Bias 17
    Noise 24
    Variability Comparison 25
    Correlation and Related Measures 26
    Trends Resulting from Changing Emissions 28
    Graphs 29
    Analysis of Subgroups of the Hourly Concentrations 30
    Ways of Pairing Daily Maximum Concentrations 31

III. EXAMPLE PERFORMANCE EVALUATION 34
  A. MODELS AND DATA BASE 35
    Data Base Used 36
    Meteorological Conditions 37

  B. GENERAL DATA SET EVALUATION 39

  C. FIRST LEVEL PAIRED COMPARISONS ON HOURLY DATA 41
    General Overview Statistics 41
    Model Performance by Hour of the Day 42

  D. SECOND-LEVEL COMPARISON: DIAGNOSING ERRORS IN HOURLY PREDICTIONS 44
    Missing the Peak in Space 45
    Point Source Influence 49
    Introduced Error 52
    Dispersion (Vertical Mixing Rate) Influence 58
    Missing the Peak in Time 62

E. COMPARISON OF DAILY MAXIMUM CONCENTRATIONS 63
Local Site Maximum for Each Day and Site, Paired by Hour . 63
Local Site Maximum for Each Day and Site, Unpaired by Hour 64
All-Station Daily Maximum, Paired by Site 66
All-Station Daily Maximum, Unpaired by Site 67
Area-wide Daily Maximum, Over Entire Modeling Region . . . 68
Regression Analysis of the Daily Maximum Pairings 69
Evaluation of the Models Based on Daily Maxima 70

F. EMISSIONS CHANGE COMPARISON 72
Definition of an Emissions Change Comparison 74
Development of the Earlier Emissions Inventory 75
Choice of Models and Days for the Emissions Comparison . . 77
Estimation of the Change in Observed Maxima 78
Results 81

G. PERFORMANCE EVALUATION CONCLUSIONS 89
Performance Character of the Model 90
Insights on the Regulatory Use of the Model 94

IV. IMPLICATIONS FOR PERFORMANCE MEASUREMENT 98
A. CONCLUSIONS OF THE USE OF STATISTICAL TECHNIQUES 98
Evaluation of the Performance Measures 98
Evaluation of Graphical Displays 102
Evaluation of the Use of Subgroups of the Hourly Concentrations. 102
Estimating the Bias in the Predicted Peak Concentration . . 104
Problems in Comparing Models on Hourly Data 106
Effects of Non-normality on Bias Comparisons 108

B. RECOMMENDATIONS 110
Recommended Performance Measures 110
The Use of Statistical Measures as Performance Standards . 113
Evaluating the Usefulness of a Model 114

REFERENCES 117

TABLES 121

FIGURES 147

I. INTRODUCTION

In the implementation of laws to protect air quality, atmospheric dispersion models have come to be a basis for establishing acceptable levels of emissions control for air pollution sources. To justify their use as regulatory tools, the models should be accurate and be used correctly. In recognition of this need, a workshop on dispersion model performance was convened jointly by the Environmental Protection Agency and the American Meteorological Society in September 1980 to recommend procedures to evaluate model performance. As a step toward improved, consistent evaluation and comparison of models, workshop participants proposed a list of performance evaluation measures. They called for testing of those measures and for further development of evaluation methods through actual tests of models (Fox, 1981).

The present study developed as a result of the AMS/EPA workshop. Three versions of the Urban Airshed Model developed by Systems Applications Inc. (SAI) for simulating the production of photochemical pollutants were tested on ozone data for Denver, Colorado. Statistical measures and graphs of model performance were used to compare and evaluate the three model versions in a complete example evaluation. Then, in turn, the usefulness of the measures themselves as evaluation tools was assessed.

This report begins with a brief discussion on the selection of evaluation criteria. It is argued that a performance evaluation, to produce relevant results, should be structured around the attributes which are important in the use of the model. Models discussed in this report are intended for regulatory use, and model attributes are selected based on an analysis of that use. In the second section, general requirements for the use of statistical measures and statistical tests are discussed. Then some measures, associated statistical tests, and graphs are described, chosen on the basis of the model attributes selected for the example performance evaluation. The third section describes the example performance evaluation for Denver, including the use of performance measures and detailed sensitivity studies that were needed to diagnose errors in the modeling. Conclusions are drawn about the performance of the model on Denver data and suggestions for further research and improvements are made. In the final section, the performance measures themselves are evaluated, and their appropriateness as performance standards is discussed.

A. SELECTION OF EVALUATION CRITERIA

A model performance evaluation needs to be structured around the intended application of the model and the objectives of the evaluation. This will determine the scope and the methods which are appropriate. Such a statement seems almost obvious, yet it was found, both in the example evaluation described here and in earlier evaluations, that important model characteristics can be easily overlooked. Initial attention to structuring the evaluation can reduce the awkward need to add new analyses and data at a later stage.

Evaluation of a Model for Scientific or Regulatory Purposes

This study focuses on evaluation of air quality models used for regulatory purposes. Evaluation of a model intended for regulatory use requires a somewhat different orientation than the usual scientific approach to model testing. A "scientific" evaluation generally focuses on determining how well the model results mimic the observed behavior of pollutant concentrations. It establishes the accuracy of the model in duplicating the magnitude, location, and timing of concentrations under selected conditions. The scientific evaluation is generally designed to explore the details of a model's performance, to determine that the model makes the right predictions for the right reasons and to identify errors as a step toward improving the model.

In contrast, a "regulatory" evaluation needs to determine how well the model provides the results necessary for decision making. Thus the purposes and methods of the model's application will define what must be covered by the regulatory evaluation. The structuring should take into account how the model is used *in practice* if meaningful results are to be obtained. This may affect the selection of both the data base and the performance measures to be used in an evaluation. Of course, in evaluating a model for regulatory use a determination of the accuracy of model predictions is needed, but the emphasis may be different than in a scientific evaluation. It will be necessary to ask how the model predictions will be used, how the effects of errors can be minimized, how much error can be tolerated, and whether there are some errors which cannot be tolerated.

Requirements Due to the Type of Model

A time-dependent photochemical model is required for predicting ozone concentrations because of the complexity of the chemical processes involved in ozone production. Ozone is a secondary pollutant. It is not emitted directly but is produced by chemical reactions between several other pollutants, primarily hydrocarbons and oxides of nitrogen (NO_x). Therefore ozone concentrations depend on (1) emissions of several pollutants; (2) the rate of chemical reactions, which depends on the intensity of the sunlight; and (3) the amount of mixing of the pollutants, which depends on meteorological factors such as wind speed. The relationship between precursor emissions and ozone is nonlinear, thus simple rollback models are inappropriate.

In any time-dependent air quality model the entire day's pattern of pollutant production contributes to the daily maximum. Therefore the model should be able to reproduce the entire hour-by-hour diurnal pattern of ozone concentrations on a given day. Understanding of errors in the peak prediction will require knowledge of the full day's predictions. This implies that both daily peak predictions and hourly predictions should be evaluated.

Requirements Based on Use of the Model

The Urban Airshed Model is used in long-term air quality management required under the federal Clean Air Act. Development of State Implementation Plans (SIPs) under the Clean Air Act necessitates prediction of a relative change in pollutant concentrations due to a change in emissions for a worst-case day. Pollution control strategies must be found which will reduce emissions enough from today's levels to achieve a required reduction in ozone concentrations by a certain date. Typically the model is used to simulate only a few worst-case high ozone days selected from historical data. Then different levels of future emissions are assumed, corresponding to different control strategies, and the model is used to predict the pollutant concentrations that would result from the changes in emissions alone, keeping the meteorological input to the model the same.

Knowledge of this method of application was important in determining the objectives of the evaluation, and the data and measures that would be needed. First, use of the model is confined to high ozone days and is focused on prediction of a daily maximum ozone concentration. Thus accuracy in the magnitude of the peak prediction on high ozone days will be of primary importance. Furthermore, in practice the model is used to simulate only a few historical days, thus it must be able to replicate concentrations under meteorological conditions specific to a given day.

Second, it is important that the model accurately predict changes in ozone due to emission changes. The variation in concentrations in a given year is primarily due to differences in meteorology, rather than differences in emissions. Major changes in emissions from the automobile fleet occur over a perod of years. Therefore the data base for model evaluation should include points in time which are sufficiently spaced that emissions changes will have occurred, and a measure must be found to compare observed and predicted trends in concentrations. Past evaluations of the Urban Airshed Model have considered variations in meteorology but have not looked systematically at changes in pollutant emissions (Hayes, 1979; Cole, 1982a and 1982b).

Model Attributes Selected for the Denver Performance Evaluation

The above considerations led to selection of the following attributes as the focus of this evaluation of model performance for Denver.

1) Accuracy of the magnitude of the peak prediction for each day.
2) Accuracy of hourly predictions, including the daily pattern of the predictions.
3) Accuracy of predicted trends in peak concentrations resulting from emission changes over a period of several years.

The first and third attributes are of primary operational importance and would be given the most weight in the evaluation or selection of a model for regulatory use. The second is important as an aid in interpreting the results, for establishing confidence in the models and for diagnosing errors. These three attributes determine what performance measures and what types of data will be needed to make relevant judgments about the model.

B. PRACTICAL LIMITATIONS ON PERFORMANCE EVALUATION

In practice, limited access to data may prevent the complete evaluation of all desirable model attributes. For example, in the Denver evaluation, the available measurement network had only five monitoring stations, not enough to adequately investigate the spatial distribution of the ozone concentrations. Attention given to structuring the evaluation, by listing the important attributes for a given application, increases the chance that the most important aspects will be covered and promotes an awareness of the ways in which the evaluation is incomplete. Such awareness is likely to be important when the results of the evaluation are to be interpreted.

II. DISCUSSION OF THE PERFORMANCE MEASURES

This discussion of statistical measures is focused on evaluation of time-dependent urban models. However, many of the statistical considerations will apply to evaluation of other models as well. The first part of this section discusses several statistical issues from a theoretical point of view. These issues are important in the selection of a data base and in the use of formal statistical tests and confidence intervals. The second part of this section explicitly assembles a set of performance measures that is most appropriate for evaluation of an urban airshed model in regulatory use. The value of certain time and space pairings is also discussed.

A. BASIC ASSUMPTIONS OF STATISTICAL TESTING

Most standard statistical tests require that the sample be selected randomly and independently and that the population be normally distributed. Each of these requirements presents special problems in our analysis and will be addressed separately. First, randomness is discussed as a problem in defining what population a sample actually represents, since model-testing samples are invariably small with limitations outside of the control of the researcher. Second, the amount of deviation from normality is examined and its effect discussed. Third, the lack of independence due to autocorrelation is discussed and adjustments to the statistical tests are computed. Finally, other sources of dependence are delineated.

Choice of a Data Sample for Model Evaluation

Selection of sample data to be used for testing the model is an important element of model evaluation which was not discussed in the AMS/EPA workshop report. Bias in the sample could easily bias the comparison of two models, for example.

If it is necessary to confine testing to a small number of days, we need to define the type of days for which it is important to evaluate the model. Then every effort should be made to assure that the days chosen are statistically random in all other respects. If the days are not representative then we should specify the limited population of days which the study describes. The validity of the model to describe a different population of days must rest on physical arguments and should not be assumed.

If models are to be compared, they should be tested on the same data. Serious biases in the comparison may be introduced by indiscriminately comparing confidence intervals derived from two models using different data sets. If two models are tested on data sets from different urban locations or years, the comparison may be biased by systematic differences between the two population data sets. Such systematic differences may be due to differences in meteorology, differences in background fluxes, or differences in HC-to-NO_x ratios.

Normality of the Data

Many researchers have found that a full year's pollutant data should be transformed to approximate a normal distribution, using logarithmic or exponential data transformations (e.g., Larsen, 1971). Confining our population to only the high ozone days seriously changes the shape of the distribution, however.

Hourly concentrations on high ozone days cover the entire range of the year's concentrations, from near zero in the early morning each day to the annual maximum on one afternoon. Histograms of the set of all high ozone ($\geq$.10 ppm) summer weekdays, May-September 1975-80, are shown in Figure 1, for each of the five Denver area monitoring sites. Most of the distributions are bimodal. The spike at the low end of four of the histograms results from low early morning concentrations at those sites. The mean and median are nearly identical in all five distributions, an indication that they do not conform to the typical skewed shape of a log normal distribution. Indeed, attempts to transform the data led to even greater deviations from normality. As a result, no data transformation was used for the remainder of the analysis.

The set of daily maximum concentrations on high ozone days actually represents the upper tail chopped off of a distribution which may, perhaps, be log normal. This upper tail deviates greatly from normality, therefore only those statistical tests whose probability levels are affected little by deviations from normality will apply.

Effect of nonnormality on bias estimates. Bias is estimated from the mean of the model residuals C_o-C_p. In a paired comparison, it is the residuals, not the original concentrations, that must be normally distributed. Furthermore, the Central Limit Theorem tells us that if the sample size is reasonably large (greater than 50 observations) then a distribution of sample means is approximately normal, even if the original data was quite nonnormal. As a result, the bias estimates from large samples can generally be assumed to be normally distributed, regardless of whether the concentrations are normally distributed, and standard t-test procedures for establishing confidence intervals on the bias estimate are appropriate (with any necessary adjustments for autocorrelation). However, if sample sizes are small, confidence intervals based on Student's t will not accurately reflect the specified probability levels. For small samples, tests which do not involve normality assumptions should be tried in determining the significance of the bias.

Effect of nonnormality on variability and noise estimates. Here, lack of normality can be a more serious problem. Both the F-test to compare variances, and the chi-square-based confidence interval on a variance estimate, require that the concentrations be normally distributed.

Empirical studies, however, have shown that departures from normality have only minor effects on the confidence levels associated with the F-test, particularly if the parent populations have similarly-shaped distributions (Myers, 1979). Markedly skewed distributions have been shown to produce an overestimate of the alpha level, making the F-test more conservative. It should be remembered, however, that use of the F-test on skewed data will be less sensitive, increasing the chance of not detecting real differences (Type II error).

Lack of Independence

Autocorrelation effects. Autocorrelation in a time series is the tendency of an effect to carry over from one element of the series to the next. Many statistical techniques assume that data points are random and independent. But in an autocorrelated time series, successive data points are not independent. Indeed, oxidant and carbon monoxide concentrations have been shown to be highly correlated from one hour to the next and, to a lesser extent, from one day to the next. This lack of independence greatly reduces the precision of estimates of the population mean and variance.

Autocorrelation in the concentrations. Our population in the Denver investigation was the set of hourly ozone concentrations on summer weekdays in which the maximum concentration exceeded 100 ppb. To avoid daily autocorrelation, successive days were deliberately avoided in selecting a sample (at the expense of some loss of randomness). Hourly autocorrelation could not be avoided, therefore its effect on precision must be assessed.

The large sample of all high ozone summer weekdays, 1975-80, was used to estimate the amount of hour-to-hour autocorrelation within a day. Hourly means were subtracted from each hourly concentration to remove the effect of the diurnal pattern, creating a stationary time series of deviations from the mean diurnal pattern. Autocorrelations were then computed for lags of 1 to 6 hours within a single day.

The average autocorrelation function over the 5 monitoring sites was .69, .38, .21, .11, .08, .08 for lags 1 to 6. When each of the 5 sites was considered separately, 4 sites followed a pattern quite similar to the average. This autocorrelation function is fairly typical of a first-order autoregressive process. Such a process has the form

$$y_t = \phi \, y_{t-1} + a_t$$

and for lag k, the autocorrelation function is

$$r_k = \phi^k.$$

Thus the first few autocorrelations can be used to estimate ϕ, i.e., $r_1, \sqrt{r_2}, \sqrt[3]{r_3}$ are all estimates of ϕ. We have taken the average of these as our estimate, obtaining $\phi = .63$.

One site, Highland, showed considerably higher autocorrelation than the others, with autocorrelations of .86, .65, .49, .37, .25, .20 for lags 1 to 6. Thus we have assumed a first order autoregressive process with an estimated $\phi = .82$ at Highland.

Hirtzel and Quon (1981) have derived the equivalent number of independent measures for n autocorrelated data points in a first-order autoregressive process with autocorrelation ϕ at lag 1.

$$n_e = \frac{n^2(1-\phi)^2}{n(1-\phi^2) - 2\phi(1-\phi^n)}$$

Within each day our data consist of $n = 12$ hourly concentrations. At the 4 sites for which $r_1 = .63$, the effective number of independent observations for a day is computed to be $n_e = 3.3$. At Highland, where $r_1 = .85$, the effective independent n is only $n_e = 1.9$. Thus the precision of estimates of $\bar{C}_o$ and S_o is equivalent to that obtained with only 2 or 3 independent measurements per day.

Autocorrelation in the Residuals. In an ideal model the residuals would not be autocorrelated even when the concentrations are. The residuals of the three models examined here all exhibit both a strong diurnal pattern and considerable autocorrelation, however.

As a result of the high autocorrelation, then, the standard error of a single day's bias estimate is not $\sqrt{\frac{S_d^2}{12}}$ as would be the case with 12 independent measurements. Instead, under a first order autoregressive process the standard error of the bias is $\sqrt{\frac{S_d^2}{n_e}}$. The confidence interval on the bias estimate will be much wider, and the precision of the estimate much lower, than if successive residuals had been independent. Whenever all of the hourly concentrations or residuals are included in a statistical analysis, the standard errors and the degrees of freedom used to compute confidence intervals must be adjusted accordingly.

The set of daily maximum concentrations will not be autocorrelated because we have excluded successive days from the sample. For the same reason, statistics can be computed for each hour separately with no autocorrelation effect.

Other Sources of Dependence

One other source of dependence between data points should be mentioned, although we have no way of adjusting for it directly. There is likely to be some correlation between concentrations occurring at the same time at different sites. Thus when statistics are averaged over all of the sites, measurements from the separate sites will not be entirely independent. Standard errors are likely, therefore, to be underestimated and any confidence intervals will be narrower than they should be.

B. PERFORMANCE MEASURES

The introduction to this report emphasized the need to specify the attributes on which a particular model is to be evaluated. Those attributes provide a basis for choosing performance measures which are relevant to the model and its application. Although an extensive list of performance measures was recommended by the AMS Workshop, some measures from the Workshop list may not be appropriate for a particular model, and additional measures may be needed. For example, in examining how the Urban Airshed Model is used in practice it was noted that only a few days are simulated, therefore the model must replicate concentrations under meteorological conditions specific to a given day. Thus matching the observed and predicted concentrations on a given day is necessary. It would be inappropriate to use the unpaired t-tests and comparisons of frequency distributions that were recommended by the AMS Workshop for use in evaluating point source models. Furthermore, the issue of emissions change was not addressed by the AMS Workshop, therefore an additional measure was required beyond the workshop list.

The measures discussed below were selected for the Denver example evaluation, based on the list of important attributes of model performance which was developed in the introduction. The measures were related to the attributes of model performance as follows:

(1) Accuracy of peak predictions: Bias, gross error, noise, variability, correlation, linear regression.

(2) Accuracy of hourly predictions: Bias, gross error, noise, variability, correlation.

(3) Accuracy of trends in peak concentrations resulting from emission changes: Difference between observed and predicted linear trends in ambient concentrations.

One potentially useful measure, spatial correlation, was not used because the data needed were not available.

Gross Error

Two measures of gross error were suggested at the AMS Workshop.

Mean square error: $$MSE = \frac{\Sigma(C_o - C_p)^2}{n}$$

Absolute deviation: $$\overline{|d|} = \frac{\Sigma|C_o - C_p|}{n}$$

Both provide an overall measure of model inaccuracy. They are so similar that they are basically redundant, therefore only one would be needed as a performance measure. Which is chosen is likely to be a matter of personal preference. For each, it is desirable to obtain a low value to minimize the inaccuracy of a model.

The absolute deviation is easy to interpret and is less sensitive to outliers (more robust) than the MSE. However, it provides no basis for computing a confidence interval.

The mean square error is familiar to regression analysts because it is the quantity which is minimized in the least squares process. Because the errors are squared, greater weight is placed on the larger errors.

Although it is possible to construct a confidence interval, this is generally not done because the distribution of the MSE is not a standard one--it is a compounding of a normal distribution and a chi-square distribution.

Our preference is for the absolute deviation, because of its interpretability and robustness. However, other, more specific, measures are likely to be more useful than either of the gross error measures. From a theoretical point of view, for a performance evaluation it should be more informative to consider separately the two components of the gross error: bias and noise.

Bias

The estimated bias in a model is the mean of the differences between observed and predicted values

$$\bar{d} = \bar{C}_o - \bar{C}_p \ .$$

If the bias is significantly different from zero, it indicates a systematic tendency of the model to underpredict (if $\bar{d} > 0$) or overpredict (if $\bar{d} < 0$).

On any particular sample of test data, of course, some difference from zero is to be expected since no sample can perfectly capture all of the characteristics of the population. Similarly two models may show slightly different bias levels on a given set of sample data when that difference should be attributed to characteristics of the sample rather than real differences in bias between the models. The model bias computed from a single data sample is merely an *estimate* of the true bias in the model. Computing a confidence interval around that bias estimate indicates how much the true bias can reasonably be expected to deviate from our estimate. In a sense, the confidence interval provides a measure of the precision of the bias estimate.

Bias comparisons based on Student's t. An EPA list of statistics, recommended for use in evaluating air quality models compiled and developed in more detail by W.M. Cox from the AMS-recommended list, suggests estimating the bias in two different ways--using both a paired t and an unpaired (2-sample) t. We would argue that the paired t is sufficient for our purposes and that the unpaired t will offer no additional useful information; therefore the difference between the two methods should be discussed.

The estimated bias is the same under both methods, that is,

$$\bar{d} = (\overline{C_o - C_p}) = \bar{C}_o - \bar{C}_p$$

However, the confidence intervals will be different, and the interpretation of "bias" will be somewhat different. In order to make our bias estimate as precise as possible we would like to obtain a confidence interval that is as narrow as possible.

Statistics on the paired concentrations provide a more stringent test of the model because they require some match (in time or space) between the observed and predicted concentrations. A completely *unpaired* test looks at the observations and the predictions as two independent data sets, with no matching whatsoever and simply asks, "Could these two independent samples have come from the same population?" and, if not, "What is the difference between the means of the two populations from which they came?"

The paired test assumes that there may be some correlation between observed and predicted values, and accounts for it. The unpaired test assumes that there is no correlation between the observed and predicted values, therefore it does not account for any correlation. Specifically, under the paired test the standard error of the bias estimate is

$$s_{\bar{d}} = \sqrt{\frac{s_{C_o}^2 + s_{C_p}^2 - 2\ \mathrm{cov}\ (C_o, C_p)}{n}}$$

with n-1 degrees of freedom, where $\mathrm{cov}(C_o, C_p)$ = the covariance between the observations C_o and the predictions C_p.

Under the unpaired t-test the standard error of the bias estimate is

$$s_{(\bar{C}_o - \bar{C}_p)} = \sqrt{\frac{s_{C_o}^2 + s_{C_p}^2}{n}}$$

with 2(n-1) degrees of freedom. (If samples are small this requires an additional assumption that the variances of the C_o and the C_p are equal, which requires that the sample variances $s_{C_o}^2$ and $s_{C_p}^2$ pass an F-test. If the variances fail this test then n-1 degrees of freedom must be used.)

When the covariance (or correlation) is zero these two standard errors are the same. If C_o and C_p are positively correlated at all, then $S_{\bar{d}}$ will be smaller and will generally produce a narrower confidence interval (unless the sample size is very small, in which case the higher degrees of freedom for the unpaired test will make for a lower critical value of t). Therefore, if $r_{C_oC_p} \geq 0$, there is no point in computing bias based on unpaired concentrations, unless sample sizes are very small and sample variances meet an assumption of equality.

If the observed and predicted concentrations are negatively correlated, the unpaired t will, indeed, produce a narrower confidence interval. But a negative correlation puts the validity of the entire model in doubt--the precision of the bias estimate assumes minor importance. Thus, as a general practice, a one-sample (paired) t should be used to establish a confidence interval for the estimated bias, and a two-sample (unpaired) t should not be required.

For comparison of more than two groups, a two-way analysis of variance is more appropriate than multiple t-tests. If many confidence intervals based on t are computed, one should be aware that the chance of making at least one Type I error increases as the number of t-tests increases. The confidence intervals are useful because they dramatize the fact that the computed biases are only estimates. In choosing a 95% confidence level, then applying it many times, it is important to recognize that 5% of the computed confidence intervals will not contain the true bias.

Bias comparisons based on the Wilcoxon Test. When samples are small and there is reason to believe that the data population is not normally distributed, confidence intervals based on Student's t may be misleading. On extremely skewed data, one test showed that a sample size of $N = 40$ was required to achieve an accurate 95% confidence interval on the sample mean, and in this case the errors were not symmetrically distributed but were confined primarily to one tail of the distribution of sample means (Barrett and Goldsmith, 1976). Thus, when it is known that data is not normally distributed, it may be worthwhile to try an alternative to the t-test.

The Wilcoxon Paired Rank Test requires no assumptions about the distribution of the data. It is based only on the rank-order of the measurements, and is quite sensitive to changes in the central tendency of a distribution. It is described as a "mean slippage" test by Pearson and Hartley (1976), who provide an excellent brief description and the required probability tables.

The procedure is as follows: Differences between the paired measurements are computed just as in a paired t-test. The differences are then rank-ordered in the order of their absolute values, i.e., in the ordering process, the sign is ignored. Finally, the sum T^- of the ranks for all of the negatively signed differences is computed. If the differences were randomly distributed about a mean of zero, one would expect that there would be similar numbers of positive and negative differences and that the sum of ranks for negative differences would be approximately equal to the sum of ranks for positive differences. Distributions of sample T^- values have been computed for sample sizes up to $N = 50$. For larger samples, the distribution of T^- is adequately approximated by a normal distribution with mean $E(T^-) = N(N + 1)/4$ and variance $\sigma^2(T^-) = N(N + 1)(2N + 1)/24$.

Example 1: Daily maxima predicted by the EPA1 model, selected from the entire grid area, compared with observed maxima for 11 days.

	Day										
	1	2	3	4	5	6	7	8	9	10	11
Observed, C_o	153	146	162	166	157	117	117	100	154	121	101
EPA1 predicted, C_p	119	113	102	129	109	93	82	85	105	142	82
Difference, d	34	33	60	37	48	24	35	15	49	-21	19
Rank of $\|d\|$	6	5	11	8	9	4	7	1	10	3	2
Sign of d	+	+	+	+	+	+	+	+	+	−	+

In this case there is only one negative difference, and its rank is 3. Hence the sum of ranks for negative differences is $T^- = 3$. Consulting a probability table for T when N = 11, we find that there is a two-tailed probability of .01 that T^- will fall outside of the interval [5, 61].

We conclude that the central tendencies of the observed and predicted distributions are significantly different, with $\bar{C}_o > \bar{C}_p$. Therefore, the bias in EPA1 predictions of the daily maximum is significantly greater than zero.

Example 2: Comparison of bias in DOT and EPA1 models in the prediction of daily maximum concentrations. Predictions for the 11 days are chosen from the full grid area.

	Day										
	1	2	3	4	5	6	7	8	9	10	11
EPA1 residual, d_1	34	33	60	37	48	24	35	15	49	-21	19
DOT residual, d_0	47	59	81	59	66	18	42	6	47	4	19
Diff. $d_1 - d_0$	-13	-26	-21	-22	-18	6	-7	9	2	-25	0
Rank of $\|d_1 - d_0\|$	5	10	7	8	6	2	3	4	1	9	--
Sign of $d_1 - d_0$	-	-	-	-	-	+	-	+	+	-	0

Note that the difference residual, $d_1 - d_0$, eliminates the observations from the comparison; only differences between model predictions are tested. Zero differences have no sign, therefore they are not assigned a rank. The sum of ranks for the negative differences is $T^- = 5 + 10 + 7 + 8 + 6 + 3 + 9 = 48$. For $N = 10$ ranked differences, there is a two-tailed probability of .05 that T^- will fall outside of the interval [8, 47]. Therefore with confidence level $\alpha = .05$, we can conclude that the bias in the EPA1 model is less than the bias in the DOT model on this data.

For comparison of more than two groups, the Friedman Rank Test may be used when the normality assumptions required by a 2-way analysis of variance are violated. This test, too, is described by Pearson and Hartley, and the test statistic is distributed approximately as χ^2.

Although the Wilcoxon test can be used to estimate confidence intervals using the methods of Hollander and Wolfe (1973), the confidence intervals used here are based on Student's t. Unfortunately, these confidence intervals may not be estimated accurately for small samples. By doing both Wilcoxon and t-tests on each sample and comparing the results, it may be possible to judge whether t-based intervals in a particular sample are too small or too large.

Use of Wilcoxon and Friedman tests will not be appropriate on autocorrelated data. The tests require that measurements be independent, and we do not know of any adjustments analogous to that of Hirtzel and Quon for the t-test. Therefore we will only be able to use these tests on data sets which do not contain successive hourly concentrations.

Noise

The estimated noise in a model is the variance of the differences

$$S_d^2 = \frac{\Sigma(d-\bar{d})^2}{n-1}$$

The standard deviation of the differences S_d (square root of the variance) is a more interpretable form of the noise measure because it is in the same units as the original data.

Bias and noise are two separate components of the gross error, as measured by the MSE, i.e.,

$$MSE = \frac{n-1}{n} S_d^2 + (\bar{d})^2 .$$

In the case when bias is zero, the gross error consists only of noise. In the (unlikely) case that noise was zero, the gross error would consist only of bias.

The effects of systematic bias in a model could be removed by simple proportional calibration procedures. Corrections for noise, however, are likely to be model-specific and may even require fundamental changes in the model. Thus it is of interest to know how much of the error is attributable to noise and therefore not controllable by simple calibration.

A confidence interval for the estimated noise can be easily established using the chi-square distribution, provided that the differences are normally distributed. The differences resulting from our three models all have skewed distributions, thus confidence intervals computed for S_d^2 would be only approximations.

It should be noted that the noise level can be expected to be quite large for completely paired hourly data, simply because predictions from state-of-the-art photochemical models tend to miss by a few hours or a few miles. Such errors are probably unavoidable. Strict pairing will continue to be useful for diagnostic purposes, but noise computed as a performance measure should probably be based on less strict pairings, such as the observed and predicted daily maxima.

Variability Comparison

A comparison of the variances of the observed and predicted concentrations can be useful for diagnosing errors in the model. If the variance in the predictions ($S_{C_p}^2$) is much smaller than the variance in the observed data ($S_{C_o}^2$), then the model is doing a poor job of picking up day-to-day fluctuations in ozone concentrations, and is probably holding too closely to an "average" diurnal pattern. The F-test can be used to determine if predicted variability is too small. If

$$F = \frac{s^2_{C_o}}{s^2_{C_p}} > F_{crit.}$$

(the critical value on the F-distribution, at the desired confidence level), then the predicted variance is significantly less than the observed variance.

If the variances are not significantly different then the model is producing an acceptable amount of day-to-day variation, but the F-test doesn't tell us whether that variation occurs at the same time or place as that in the observed data. If such matching is important then the noise measure is more appropriate than the variability comparison.

Correlation and Related Measures

The Pearson correlation coefficient should be used with some caution in model evaluation, since its results can be misleading. It measures the strength of a linear relationship between observed and predicted concentrations. Three problems will be addressed: 1) Linear correlation ignores the possibility of a curvilinear relationship. 2) A perfect linear correlation ($r=1$) could theoretically be obtained even when there are large errors in the model. 3) High correlations on hourly concentrations may be obtained merely because the model is able to duplicate an average diurnal pattern, regardless of its ability to simulate differences between days.

1) The best initial picture of the relationship between observed and predicted concentrations can be obtained from time series plots of C_o and C_p and scattergram plots of C_o against C_p. A curvilinear relationship between C_o and C_p, or other unexpected pattern, may become evident and may be useful for diagnostic purposes.

2) Computation of the correlation between C_o and C_p should be accompanied by computation of the slope and intercept of the regression line $C_o = a + bC_p$. A perfectly fitting model would produce not only a correlation of $r=1$ but also a slope $b=1$ and an intercept $a=0$. Errors in both C_o and C_p will affect the magnitude of r. In the above regression equation, errors in C_o will not produce bias in the estimate of b, but they will affect its sampling distribution (Brier, 1975). The scattergram, slope, and intercept may be even more useful than $\bar{d}$ for diagnosing bias in a model, since they may show systematic tendencies to overpredict at some concentration levels and underpredict at others.

3) It is of dubious value, given the strong diurnal pattern in ozone concentrations, to throw together all hourly concentrations and compute an overall correlation between C_o and C_p. The magnitude of the differences between the hours of the day is much greater than the variation within any given hour. Therefore this overall correlation primarily reflects the ability of the model to approximate the shape of the average diurnal pattern. High correlations, while mildly reassuring, often

indicate only that. Differences in capturing deviations from the average diurnal pattern will have relatively small impact on the magnitude of r. It could be worthwhile to remove the effect of the diurnal pattern by subtracting hourly means from the observed and predicted concentrations before computing the correlation.

Of course, the cyclic effect of a diurnal pattern does not enter into correlations of daily maximum concentrations. This accounts for the drastic reduction in r to be seen in the St. Louis ozone study (Cole, 1982a) when going from full-day hourly data to daily maximum data.

Trends Resulting from Changing Emissions

In the regulatory use of the Urban Airshed Model, accurate response to changing emissions is critically important. Changes in ozone maxima over a period of years are affected by both emissions and meteorology. To estimate trends due to emissions change, it is necessary to filter out the effects of year-to-year weather fluctuations.

In this study a linear regression of concentration versus year was chosen to estimate the annual trend, or rate of change, in peak concentrations. Regression was done on daily maximum concentrations C_y for high ozone days in year y, based on the trend model $C_y = a + ty$ where the trend t and the constant a are estimated coefficients.

Observed and predicted trends over a period of years were estimated separately, then compared. The trend predicted by the models due to emissions change was obtained by repeating the simulations of the original days in 1979-80 using a Denver emissions inventory for 1976. Meteorological conditions were unchanged. To obtain an observed ozone trend for comparison it was necessary to smooth out annual fluctuations in meteorology, therefore daily maxima concentrations from high ozone days for all six years from 1975 to 1980 were used in the linear regression.

Graphs

Graphs are an integral part of statistical modeling, at every stage from model development through model evaluation. In particular, scatterplots of the residuals of a model are the primary tools recommended by Draper and Smith (1966) for evaluation of a regression model. We suggest the following graphs.

(1) Histograms of the concentrations and of the differences, to show the shape of their frequency distributions.

(2) Plots of C_p against C_o, and of both these against time.

(3) Scatterplots of the differences (model residuals) against any relevant variable, including time, observed concentration, predicted concentration, and variables used as input to the model. The residuals should be scattered randomly within a horizontal band of even width. If their pattern is sloped, curved, or cyclic, then inadequacies in the model may be indicated.

(4) If the data fall into natural categories of some type, plots of residuals by category, or bias by category. If the residuals are normally distributed, then confidence intervals on the bias should also be plotted. If not, box plots of the residuals by

category would be a useful alternative (Kleiner and Graedel, 1980).

Because pollutant data falls into natural categories by monitoring site, by hour of observation, and by day of observation it may be useful to do plots for each of these categories separately. In our experience, plots of bias by category, rather than the residuals themselves, have provided the best visual checks on possible patterns of error in a model.

Graphs can also be very helpful in the search for specific causes of error in a model. For example, in the Denver evaluation plots of the observed and predicted spatial field of concentrations and plots of wind trajectories were found to be useful.

Analysis of Subgroups of the Hourly Concentrations

Frequently it is worthwhile to do special analyses on subgroups of a data set in order to detect characteristics unique to that subgroup which may be masked or averaged out in the full set. One AMS workshop suggestion involved separating the data into meteorological categories, for example by stability class or wind speed, and comparing model performance on different categories. Such categories should not be used with hourly observations if confidence intervals are desired, though, because it would be impossible to determine the amount of autocorrelation in the subgroups. A useful alternative is to create categories based on hour of the day, averaging over all of the days. This tends to separate the data roughly by stability class, and to show special morning and afternoon characteristics as well. By using only one hour from each day in each group, it eliminates the problem of hour-to-hour autocorrelation.

Two ways of sorting the hourly concentrations (breakdowns) were found to be useful in the Denver example evaluation:

Comparing $C_o(x,t)$ with $C_p(x,t)$ for each hour separately, averaged over all of the days. This is the most important breakdown because it addresses how well the model is reproducing the diurnal pattern. Performance measures were computed for each site and for all sites together.

Comparing $C_o(x,t)$ with $C_p(x,t)$ for each day separately, averaged over all hours in that day. This breakdown is useful as an aid to diagnosis because it can isolate days with unusual characteristics. It obscures information about the diurnal pattern, though, hence it does not contribute directly to the evaluation of performance. Again, the measures were computed for each site and for all sites together.

Ways of Pairing Daily Maximum Concentrations

The accuracy of maximum ozone predictions on a given day can be judged in several ways, depending on how the "maximum" prediction are selected.

Comparisons on the Site Maximum

A local site maximum is the observed maximum concentration on a given day at a given site. Two comparisons were tried, one requiring complete pairing in time. They were

a) $C_o^{max}(s,h)$ with $C_p^{max}(s,h)$, paired by hour of the day. The observed maximum is paired with the prediction at the same hour.

b) $C_o^{max}(s,h)$ with $C_p^{max}(s,x)$, unpaired by hour. The observed maximum is paired with the predicted maximum for that day and site, no matter what the hour.

Comparisons on the Daily Maximum

The daily maximum is the maximum concentration over all sites. Three types of comparisons were tried, representing successively less stringent pairings in space. They were

a) $C_o^{max}(s)$ with $C_p^{max}(s)$, paired by site. The predicted daily maximum at the site of the observed daily maximum.

b) $C_o^{max}(s)$ with $C_p^{max}(x)$, unpaired by site. The predicted daily maximum that was predicted at any site, whether observed there or not.

c) $C_o^{max}(s)$ with $C_p^{max}(g)$, unconstrained in space. The predicted daily maximum from any grid point in the modeled region.

If the days to be simulated are chosen randomly, then we would expect the following biases to result from these pairings. Pairing by site, (a), constrains the choice of predicted maximum, therefore it can be expected to lead to underprediction and hence to positive bias. When daily maxima are not paired by site, (b), there are equal numbers of observations and predictions to choose from, hence no bias is inherent in the pairing method. If predicted maxima are unconstrained in space, (c), then overprediction, hence negative bias, should be expected because observed maxima are limited to the monitoring sites.

In practice, however, studies of model performance are often confined to days on which high ozone concentrations have been observed at the monitoring sites. This should produce some additional tendency toward underprediction under all three pairings, because days with lower observed concentrations have been excluded.

III. EXAMPLE PERFORMANCE EVALUATION

This section presents the use of the performance measures described above in an actual performance evaluation. The goal is to discover which measures give the most information for different purposes of evaluation, and to demonstrate the complexity of interpreting the information contained in performance measure statistics. It will be shown that skill is required, and that a performance evaluation most likely cannot be performed in a routine mechanical fashion.

First, the data set as a whole will be examined, looking for anomalous behavior. Then, a first-level comparison of hourly observed and predicted ozone concentrations will be presented. Because an understanding of the sources of error is so critical, a second level comparison of hourly data is presented, showing how a combination of statistical measures and sensitivity analyses can be used to diagnose the causes of error. The experience with the performance measures on hourly concentrations is summarized before next turning to the peak concentration comparisons. Several ways of matching peak concentrations are presented, moving through successively less restrictive constraints in the pairing of data. After summarizing these comparisons, the performance evaluation is taken one important step further: the model is tested for the regulatory purpose for which it was designed--the prediction of changes in concentrations due to changes in emissions. Finally, concluding comments are made about the carrying out of a performance evaluation.

A. MODELS AND DATA BASE

This study was intended to evaluate the use of statistical measures to discriminate between models as well as to evaluate a single model. Therefore, three versions of an urban photochemical model were compared on the same data base. The basic model is the Urban Airshed Model, developed and modified by Systems Applications Inc. (SAI) for use in air quality planning work required under the federal Clean Air Act. A description of the model and its usage is given in an EPA guideline (Layland, 1980). Variations on this model have been used previously in Denver (Reynolds, 1979) and in Los Angeles (Tesche, 1981 and Reynolds, 1979), Sacramento (Reynolds, 1979), St. Louis (Cole, 1982b), Tulsa (Reynolds, 1982) and Philadelphia (Haney, 1983).

The three versions represent incremental improvements in the model. The earliest version, which will be referred to as the Department of Transportation (DOT) model, uses Carbon Bond I chemistry (Reynolds, 1979). The intermediate version, referred to as the EPA1 model, uses an improved chemical mechanism, Carbon Bond II (Whitten, 1980). The most recent version, referred to as the EPA2 model, uses Carbon Bond II chemistry and, in addition, reduces the artificial dispersion of pollutants within the model by using an improved finite differencing method (Schere, 1982).

The EPA2 version is the model currently recommended by the Environmental Protection Agency for prediction of photochemical air pollution in State Implementation Plan work required by the Clean Air Act. Our testing of the EPA2 version on Denver ozone concentrations is one segment of a broader evaluation of that model which also includes St. Louis and Philadelphia.

Data Base Used

Eleven days were simulated for the performance work. These were selected from high-ozone days in Denver in the summers of 1979 and 1980 having a peak ozone concentration of at least 100 ppb. (The maximum observed was 166 ppb.) Only weekdays were included, because an emissions inventory was available only for weekdays, but this is not a serious restriction because observed ozone patterns are similar for weekends and weekdays. The sample days, by and large, represent isolated high ozone days--single day "episodes." This is typical for Denver's highest ozone days. The sample of days was also weighted towards the days with the highest observed maximum.

Figure 2 shows the region modeled in and around Denver. The shaded areas show the contiguous metropolitan and developed regions. The five monitoring stations are also shown: Arvada, CAMP, CARIH, Highland and Welby. The modeling grid of 2 mile by 2 mile cells has been overlain for perspective.

In all cases but one, the modeling area shown in Figure 2 fully contained the predicted daily maximum. For one day, the peak was at the edge of the modeling region boundary. It is estimated that this did not affect the predicted daily maximum by more than 5%. The simulations were run from 5 a.m. to 5 p.m. (1700), the time over which photochemical production takes place. All daily maxima, predicted and observed, were contained in this time period. Because our population of days represented single-day episodes, there was no reason to carry the simulations further in time, given limited computer resources.

Meteorological Conditions

Several types of data are included here to provide background on the meteorological conditions that existed on the eleven days that were simulated. We have not made a thorough analysis of these meteorological conditions compared to meteorology on days with low ozone levels. Thus we cannot indicate which conditions are better than others as indicators of high ozone days. These data do, however, give some indication of the similarities between the days which were modeled.

Our judgment, after reviewing the performance evaluation of the modeled days and reviewing these data, is that there does not seem to be any pattern to suggest that some of the eleven days would be associated with multi-day periods of stagnation; rather, the eleven days have been judged to represent individual, single-day developments of high ozone levels. This seems to be the dominant and special character of Denver's ozone problem. The most prevalent characteristics of the days are strong heating at the surface, very low wind speeds throughout the morning until mid-day, low wind speeds through early afternoon, typical summertime mixing depths and a high pressure ridge aloft. A variety of wind patterns, not a single type, characterize these days, but the dominant pattern is one of wind flows zigzagging over Denver.

Table 1 shows the synoptic conditions of high pressure at 500 mb and at the surface for the modeled days, together with the conditions one and two days earlier and one day after. These data are from the Daily Weather Maps of NOAA. No pattern is evident between the sequence of high pressure at the surface and daily maximum ozone values. A "pattern" is evident for high pressure at 500 mb. All of the days in our data set are ones in which a high-pressure ridge moves slowly over Colorado at 500 mb, being there the day before and the day after the modeled day.

Further data related to synoptic conditions are given in Table 2. Of note is the fact that the surface temperature on almost all of the modeled days is above 90°F, a high temperature for the Denver area. Precipitation is associated with thunderstorms in the afternoon. This is evident in Table 3, showing sky cover by time of day. It is noteworthy that insolation is strong through noon. The time of maximum temperature corroborates that strong surface heating is occurring on each of the eleven days.

Strong surface heating implies that these days should have high mixing depths. One must take subsidence into account, however. Table 4 shows that only one day had an upper level inversion below 2100 m at 0500 MST and no day had evidence of an upper-level inversion at 1700 MST. This together with the fact that there were no sudden changes in upper-level dewpoint temperatures implies that subsidence is not a factor that needs to be considered on these days. There appears to be no relationship between estimated mixing depth and observed daily maximum ozone on these eleven days.

The most notable and common meteorological condition across the eleven days is the existence of low winds in the morning. This is shown in Table 5 for the wind speeds at the five ozone monitoring stations. The morning wind speeds are low and even the average wind speed for the day is low, not far from 2 m/sec.

The character of the wind flows in time are seldom simple for the eleven days. Four rough categories are sufficient to give the appropriate impression (Table 6). As shown in Table 6, there is a simple straight through wind flow on day 79180. Day 79249 had a straight through flow interrupted by a zigzag over Denver. Several days had mostly a zigzag wind flow over Denver and three days showed curved wind flows. Day 79218 was interesting because it had a smooth wind reversal over much of the urban area.

B. GENERAL DATA SET EVALUATION

Before going through an evaluation, it is worthwhile to check whether any days show anomalous behavior. This will point out unusual days for which the simulation should be checked, either because the model behavior is out of the ordinary, or because the data base contains unusually large errors. The bias and absolute deviation computed for each day separately can show average differences in performance by day. It is helpful to compute these as a percent of the mean observed concentration for the day. The daily bias and daily absolute deviation give similar information, but because we want to understand how the model is doing on the whole (looking for gross errors), the daily absolute deviation is expected to give a more complete indication. Figure 3 shows the daily bias and the daily percent bias, while Figure 4 shows the daily absolute deviation and the daily percent absolute deviation, for all three models. The differences between days are minor in all four graphs, and the daily percent absolute deviation is particularly uniform across the days. If any day showed an absolute deviation far higher than the others, or a bias that was significantly larger, it should receive special analysis. It is interesting to note that those three days on which the observed peak occurred at the Highland site are the days which have the highest daily bias.

Figure 5 (a) shows the daily absolute deviation plotted against the observed daily average concentration. We are interested in detecting any unusual relationship between a day's gross error and its average ozone level because this, too, could indicate a day which requires special attention. Again, however, no day stands out as unusual.

For a thorough check on anomalous behavior, one should use confidence intervals on the bias estimates as shown for EPA2 in Figure 5 (b). Autocorrelation in the model residuals has been taken into account in computing the confidence intervals. Residuals of all three models investigated here contain substantial autocorrelation, as shown in Table 7. The exponential decline of each autocorrelation function with increasing lag is characteristic of a first order autoregressive process. The autocorrelation in the EPA1 and EPA2 model residuals is slightly lower than in the DOT model residuals, however the differences are too small to be significant.

If the 12 hourly residuals from a single day are used together, the effective number of independent observations ranges from $n_e = 2.9$ for the DOT model to $n_e = 3.3$ for the EPA2 model. (Computed using the equation on page 12.) The estimates of ϕ show considerable variation, however, so $n_e = 3$ per day will be used in computing the standard error of the bias for all three models, in order to avoid assuming more precision than is warranted.

From the confidence intervals in Figure 5(b) it can be seen that bias estimates do not differ significantly over the 11 days. With only five monitoring stations, the small sample size makes it unlikely that differences between days would be statistically significant, therefore a day would have to be rather far out of line to be termed anomalous in the Denver example. Figures 3, 4, and 5 give consistent information about the data set, indicating that no one day exhibits anomalous behavior. We conclude that all 11 days can be used for the performance evaluation, for all three models.

With this conclusion, we now turn to the first step of the evaluation itself: the paired comparisons.

FIRST LEVEL COMPARISON: PERFORMANCE MEASUREMENTS FOR HOURLY DATA

General Overview Statistics

Performance measures for the three models on the full hourly data set are shown in Table 8. Measures are also given for each site separately to show differences in performance between sites. Autocorrelation in the differences, d, has been accounted for in establishing confidence intervals for $\bar{d}$. Because some correlation between sites is likely, the confidence interval estimates for the bias in the full data set is probably somewhat too narrow.

Several basic conclusions can be drawn from these measures. First, all three models show bias significantly greater than zero on the full data set and at the majority of sites. Thus, there is a systematic tendency to underpredict in all three models. Second, the variance in the predictions was significantly less than the variance in the observations in all three models. Third, model performance is similar for the three: bias, noise, and variability show much greater differences between sites than between models. Fourth, bias as a percent of the mean concentration $\bar{C}_o$ is particularly high at CARIH and particularly low at CAMP for all three models, with the differences bordering on statistical significance.

These general observations provide little insight into where the models may be going wrong, however. More detailed breakdown of the data is required to find which hours or days contribute most to the bias and noise.

Model Performance by Hour of the Day

One of the most important predictive capabilities of the model is its ability to simulate the diurnal cycle of ozone production. Figure 6 shows the observed and predicted diurnal patterns, averaged over our 11 sample days, for each of the five monitoring sites. The CARIH site stands out, having strong underprediction throughout the day. At the other four sites, predictions are close to or slightly higher than observed values in the morning, with substantial underprediction at the peak. This is a pattern which was also observed in an evaluation of the EPA2 version of the model on St. Louis data (Cole, 1982b). Only at CAMP do the afternoon predictions come close to observed values.

Figure 7 shows the hourly bias, averaged over the 11 days, for each of the three models for each of the five monitoring sites. It is apparent that all of the models show large bias in the midday hours, and that all of the models are basically alike in their hourly bias pattern. The bias is largest from 11:00 a.m. on at all of the stations. As noted earlier, the bias pattern at each station is different, both in shape and in the timing of the maximum bias.

Figure 8 shows EPA2 hourly bias estimates for each of the five sites, with the associated 95% confidence intervals based on student's t. None of the sets of hourly residuals differed significantly from a normal distribution under the Kolmogorov-Smirnov test, even with a significance level of $\alpha = .20$. Therefore t- and F-tests were assumed to be appropriate for this data. However, to compare the t-test with the Wilcoxon Paired Rank Test, the Wilcoxon test was also used to determine whether hourly EPA2 biases were significantly different from zero with $\alpha = .05$. For the 60 biases tested (12 hours x 5 sites, each with n = 9 to 11), the Wilcoxon and t-tests disagreed only twice, and in both cases of disagreement the confidence level was quite close to the .95 borderline. This proportion of disagreement (.033) is quite compatible with a significance level of .05. When all sites were analyzed together, with approximately 55 measurements for each hour, the Wilcoxon and t-tests agreed on significance or non-significance of the bias for every hour.

Figure 9 shows the hourly bias and the hourly percent bias for EPA2 with the 95% confidence intervals for all of the sites averaged together. With the confidence intervals, one can see that the bias is statistically significant from 10:00 a.m. on.

To give a general comparison of the models by hour, Table 9 shows the bias and noise for each hour, averaged over all of the sample days and sites. The differences between the models are small at any given hour in noise as well as bias. The bias peaks around noon, as do the observed ozone concentrations, declining during the afternoon. The noise, on the other hand, continues to increase until 3 p.m. (hour 15). Thus the errors in the models decrease in the afternoon on the average, but there continue to be high errors in the afternoon in some cases. Looking at each site separately, the high noise in the afternoon is most characteristic of Highland. Extremely high noise at Highland for hours 14 and 15 shows that errors differ greatly from day to day there, in those hours. This indicates the need for a special comparison of daily data at that site.

Daily bias, averaged over all hours and sites, was checked above (Figure 6), looking for anomalous days. To look instead for unusual model performance by day at a given site, Figure 10 shows the daily bias, averaged over all hours, for each station. One immediately notes, in this figure, that the daily bias is quite different across the stations. Furthermore, at Highland three days stand out with high bias while the others have near-zero bias.

Different patterns at different stations, both in the hourly and in the daily bias, are clues that many sources of error have been intertwined within our average estimates of bias and noise.

D. SECOND-LEVEL COMPARISON: DIAGNOSING ERRORS IN HOURLY PREDICTIONS

This evaluation is meant to examine the adequacy of the models from a regulatory perspective. Clearly the above analysis has not provided enough information to make that assessment. Systematic errors have been found but their causes, and hence their impact on regulation, have not been identified. Some errors will be tolerated for regulatory applications, others will not. It is imperative to try to diagnose the causes of hourly bias before one can truly begin to assess the adequacy of the model.

Several causes of error will be pointed out in this section. The errors were discovered and/or analyzed by using graphs, statistics on subgroups of the data, and sensitivity studies involving controlled changes in the model or its inputs. The causes of error discussed are missing the peak in space, point source influences, introduced error, dispersion influences, and missing the peak in time.

Missing the Peak in Space

Most models can be expected to place the peak in the wrong position in space, because trajectory errors are introduced in the wind field when hourly averages are used, because the monitoring instruments and sites are not perfect, because data has to be interpolated and extrapolated, and because the wind observation network is normally too coarse to resolve the necessary structure in the wind field. The airshed models seem to exhibit both errors in direction and errors in distance when missing the peak in space.

The situation at Highland is an excellent, clear example in which error in the spatial location of the predicted peak contributed a major share of the bias and noise. From Figure 10 it is clear that three days: Days 180, 218, and 249, have unusually high daily bias compared to the other days. These three days also have much higher than average daily observed ozone, as indicated below.

Day	Julian Date	Daily $\bar{C}_o$ at Highland
1	79180	90.9*
2	79193	60.8
3	79208	60.8
4	79218	92.5*
5	79249	85.4*
6	80170	58.8
7	80177	40.0
8	80191	52.1
9	80204	48.7
10	80207	45.2
11	80219	42.2
	Average All Days:	61.8

The models did not make correspondingly high predictions at Highland on these three days; rather, the predictions are similar to those on the other eight days. A check of the isopleth maps of predicted ozone concentration shows that on all three days the peak ozone cloud came near Highland, but missed the monitoring site on every occasion (See Figures 11, 12 and 13). In addition, the predicted ozone peak was earlier than the observed peak on day 79180, on which the wind simply blew southward across Denver. The predicted peak was late on day 79218, on which there was a wind reversal at mid-day over much of Denver. The predicted peak was "on time" on day 79249, on which the wind zigged a bit over Denver before continuing southward. Thus peaks were predicted nearby and on trajectories consistent with a peak being observed at Highland, but they clearly missed the monitoring station. The impact of missing the peak in space (and time) is assessed in Table 10. A significant portion of the difference between observed and predicted ozone peaks on these three days can be explained by the predicted peak having missed the monitoring site. At least half of the difference on these days, however, is associated with underprediction of the peaks by the model.

Several bias and noise measures provided a clue to this problem. In particular, the hourly noise jumped by nearly a factor of two for the hours of 1400 and 1500 (see Table 11). These are the hours when the observed peaks occurred at Highland for these three days. Table 11 also shows the change in the hourly bias and noise in EPA2 when the three days are removed and the measures re-computed. The high bias in the afternoon has essentially disappeared, being replaced by more random behavior, as one would hope to see. Thus, it appears reasonable to state that most of the bias and noise between 1200 and 1600 at Highland can be attributed to the model's behavior on those 3 days. This problem of missing the peaks in space also accounts for low variability in the model predictions at this site, since observed variability is also low in the eight-day subset which contains no peaks.

The Highland example shows that it is possible for statistics on subgroups to aid in pinpointing clear problem days. Problems can be isolated if not every day has some problem or another. While the statistics could pinpoint days contributing most of the bias and noise, they could not go to the next step and actually assess which errors were contributing to the bias and noise on those problem days. The isopleths for each hour of each day, and knowledge of the wind trajectories on a case-by-case basis, were necessary to assess the errors.

Only one of the other four stations, Arvada, exhibits some of the characteristics that seem to be associated with a day and site in which the peak has been missed in space: unusually high daily bias, unusually high observed peak ozone, and unusually large jumps in hourly noise. Arvada has four days with high bias relative to the other days. (There was one day with a large negative bias that was not large enough to be a candidate. See Figure 10.) Three of these days had a large observed ozone average, days 79180, 79193, and 80204. But Arvada does not evidence any large jumps in the noise during peak hours, therefore missing the peak in space is probably not a major contributing source of error at this site. The hourly bias and noise in EPA2 for Arvada, with and without the days 79180, 79193, and 80204, is given in Table 12. There is some improvement in the bias for the hours 1200, 1300, and 1400, but nothing as dramatic as with Highland. The hourly noise has hardly changed. As with Highland, the early morning negative bias became more pronounced for the hours 600, 700, 800, and 900.

Elimination of the days which had both high observed concentrations and high bias did not produce much change in the error measurements (bias and noise) at Arvada. We conclude that sources of error other than missing the peak in space should be found at Arvada. For example, from examination of the observed ozone concentrations in Figure 11, it is clear that on Day 180 the ozone peak is supposed to be in the vicinity of Highland, as predicted, and not at Arvada. But there should still be some ozone remaining over central Denver, upwind of the peak. Thus day 79180 does not so much represent a problem at Arvada of missing the peak in space, but rather, of missing the residual or left-over ozone when the peak is to the south.

Isopleths of ozone concentrations for days 79193 and 80204 are shown in Figures 14 and 15. Figure 14 indicates that some of the bias at Arvada is due to the model missing the peak in space. Figure 15 also suggests the peak is missed in space. Wind trajectory analysis on the predicted peak of day 80204 indicates, however, that the predicted peak may be an artifact of a 2 hour slow-down and 180° reversal in the wind field. Additionally, the vertical mixing sensitivity study discussed later finds a predicted peak at Arvada and CARIH on day 80204. Thus causes other than missing the peak in space contribute to the bias at Arvada.

In summary, three indicators simultaneously giving unusual answers--unusually large daily bias, high observed ozone on those same days, and large changes in hourly noise at peak hours--seem to be good discriminators in looking for days in which missing the peak in space is one major source of error. However, to get beyond that general identification to the level of detail necessary to support a regulatory analysis, one must use graphs and a case-by-case examination of the results.

Point Source Influence

There are still unidentified sources of bias in the Urban Airshed Model predictions for Arvada and the interior station, CARIH. As Figure 10 indicates, CARIH has a serious bias problem on almost every day, and Arvada on some days. The problem in the daily maximum predictions at CARIH is shown in Table 13. One possible cause is that the point source nitrogen oxide emissions are being mixed too rapidly to the ground (vertically) and too rapidly horizontally, effectively depressing the ozone production. Thus, this potential source of hourly bias will be investigated next.

A sensitivity study was carried out on six of the days. For this sensitivity study the point sources were removed from the input data for the EPA2 model and the day was resimulated. The days were picked to span the range of concentrations observed in our 11 sample days at CARIH, because we wanted to find out whether CARIH was influenced by the point sources. The six days also have wind flows that zigzag or curve over central Denver, near Arvada and CARIH. Thus they should represent well the influence of point sources on the prediction at these two sites.

The hourly biases which resulted from the sensitivity study are given in Table 14 for each of the stations. It is not surprising that there was no change at Highland, because there are no point sources near that site. There was also essentially no change at Arvada and Welby. There was some reduction of bias at 1000 and 1100 at CAMP, little change in the rest of the hours. At CARIH, there was some reduction of bias at 900 and 1000, but absolutely no improvement was given to the extreme bias shown at 1200. It is interesting to note that the bias at 1200 is much larger for these six days than for the 11 days on the average.

Change is more apparent when looking at the maximum ozone predicted at three sites, Arvada, CAMP and CARIH, as shown in Table 15. Arvada shows a slight but uniform improvement in the maximum predictions. CARIH and CAMP are more mixed, the change being uneven across cases, but day 79193 shows clear improvement in the model predictions. Isopleths of predicted ozone on different hours for day 79193 are shown in Figures 16, 17, 18 and 19, showing the isopleths for the case with point sources and comparisons without point sources.

On day 79193, the point sources produced a number of effects. First, they depressed the peak concentrations that were predicted (see Figure 17). Second, they cut the ozone peak apart, reducing the predicted spatial extent of the ozone cloud (compare Figures 18 and 19). Third, they caused stationary "holes" to appear in the predicted pattern of ozone concentrations (see Figure 16). These impacts on the predicted ozone are important, but they do not explain CARIH's problem. Day 79208 shows an even larger reduction (Figure 22) in the spatial extent of the ozone peak as 79193.

Day 80170 has the opposite effect on CARIH from Day 79193 in Table 15. Isopleths diagrams for this day are shown in Figures 20 and 21. The isopleths show CARIH to be in a saddle and the saddle has become a bit lower when the point sources are removed. But the height of the maximum predicted peaks and their spatial extent are not changed much at all, though the hour at which the highest peak occurs did change. As before, some "holes" in the ozone disappear.

None of the analyses of the point source influence show any cause to associate the bias at CARIH or Arvada with some unusual behavior associated with the dispersion of point sources. In other words, no significant fraction of the hourly biases that we are trying to explain can be attributed to point source effects.

The sensitivity study did show the point sources were having other important influences, however. Figure 22 compares the predictions of EPA1, EPA2 without point sources, and EPA2 with point sources for day 79208. It suggests that for some days elimination of the artificial diffusion in the change from EPA1 to EPA2, while improving the peak predictions of EPA2, made its results more sensitive to the point source emissions. In addition the comparison of contour plots from EPA1 and EPA2 showed other stationary holes in the ozone were more apparent in EPA2 than in EPA1. These other holes behave similarly to those induced by the point sources, which suggests there may be sources in the area source emissions inventory that act similarly to the elevated point sources.

Introduced Error

By "introduced errors" we mean errors which can be directly attributed to one of the inputs to the model. One case is addressed in detail in this report, hourly bias introduced through the setting of background concentrations. An understanding through sensitivity studies of how the model responds to these inputs was established before the concomitant pattern in the hourly bias was recognized. The effect of two other inputs to the model, the wind fields and the emissions inventory, will also be discussed briefly.

Other potential sources of introduced error were not examined. A known error is the use of a surface-based photolytic rate constant for NO_2. Because ultraviolet radiation, which induces NO_2 dissociation, increases with increasing altitude, higher rate constants are expected above the surface (Demerjian, 1980). This has a direct impact on ozone concentrations. For Denver, this effect could possibly account for a 10 percent underprediction of the peak ozone concentration. Other sources of error, such as the emission inventory, could easily contribute errors of this magnitude as well.

A sensitivity study was conducted to test the model response to changes in the value of background ozone. Seven of the 11 days were picked at random for this test. EPA1 was used for the test due to computer resource constraints. For the low background 20 ppb was used; for the high background 90 ppb was used. This variation was centered at 55 ppb and the average background for the seven days was 50.7 ppb. Table 16 summarizes the substantial effect on the daily maximum predicted ozone. Similar effects are seen at the individual stations, as for example on day 79218 in Figure 25. An examination of the isopleths shows that no peak was changed as to the prediction of its location by an increase in the background ozone from the normal value. See for example Figures 26, 27, 28, and 29. On three of the days (80170, 80191, and 80219) there is an apparent shift in time of the daily maximum. For the high ozone cases shown, this shift in time is caused by a change in the relative importance of two different peaks in the modeled region. For example, on day 80170, the primary peak has its maximum at 1500, measuring 92.7 ppb (Figure 28), while the secondary peak has its maximum at 1700, measuring 90.2 ppb (Figure 29). For the high ozone background simulation, the original primary peak still has its maximum at 1500, measuring 124 ppb, while the original secondary peak still has its maximum at 1700, but now measures 138 ppb, becoming the primary peak. The same behavior explains the time change of the daily maximum for days 80191 and 80219. All of the other days only had a single peak predicted in the modeled airshed; hence there was no change in the timing or location of the peak. (Day 79180's change in time should not be taken seriously because the peak is right at the edge of the modeling region.)

The simplified method used to set background ozone concentrations for the Denver model runs introduced additional bias in the early morning predictions. For each day a constant value for background ozone was used and the vertical profile of background ozone was taken to be uniform throughout the day. Nighttime chemical reactions "scavenge" ozone at the surface; thus, in the morning the amount of ozone at the surface is reduced and there is an increase in ozone concentration with height, returning to background levels. This scavenging was not taken into account for the eleven days.

Tracking background ozone through the use of monitoring data during the day was not considered to be feasible because Highland is the only station that is remotely rural, the other stations are interior to the metropolitan area (Arvada, which is upwind all day on day 79180 shows significant ozone production.) Highland was truly upwind of Denver on only one day out of the eleven.

The sensitivity studies demonstrated that stations furthest from the center of the urban area, in other words, Highland and Arvada, were most quickly affected by and responsive to the level of background ozone. The interior stations were less affected and also affected later in time. For Highland and Arvada, the background ozone value had a strong influence on the predictions as early as 700 making them the best candidates for this analysis. For all of the stations, the 1700 prediction was almost completely determined by the background value.

The previous analysis of missing the peak in space at Highland showed that removal of three days with high daily bias left eight days that had only small errors at Highland associated with production of ozone during the day. Such was not the case for Arvada. Highland also showed low variability in observed ozone concentrations on the eight days. Thus Highland is the only station for which a check might be valid on early morning bias resulting from the settings of background ozone used in the model runs. If Highland is actually a fairly good station to use as an indicator of the background ozone in the morning around the Denver area, then the existence of statistically significant bias at Highland in the early morning, as shown in Table 11, indicates that the decision to keep the background ozone constant throughout the day introduced additional bias in the early morning, an overprediction on the order of 10 ppb at 0700 and decreasing to zero by 1000 or 1100.

Results for the eight days in Table 11 suggest that the "detection limit" of the bias statistic was around 25%; that is, the bias needed to be larger than 20-25% of the observed concentration to be statistically significant at the 95% confidence level. One must remember that a sample of eight will not provide a very precise bias estimate. Thus while subgroupings of data can pinpoint different sources of error, it will probably seldom occur that those individual errors will be shown to be statistically significant, unless the model is performing very poorly.

The analysis discussed in this section indicates that either the background was set too high for the early- to mid-morning period or that something else is not being properly accounted for in the model. Given known scavenging of ozone at the surface, we believe the latter explanation. The afternoon background, on the other hand, seems to have been set about right, but it is difficult to tell. There is good correspondence between the observed late afternoon ozone for the eight days of Table 11 and the average predicted value for Highland: 57.3 ppb vs. 59.1 ppb, respectively, but the bias at 1600 and 1700 in Table 11 indicates something is still not quite right.

Further sensitivity analysis examined the effect of background hydrocarbon concentrations on ozone predictions of the EPA1 model. No attempt was made, however, to assess the effect of errors introduced via the background hydrocarbon concentrations input to the model on the bias or gross error in the model predictions. The original simulations assumed a constant vertical profile. The Tulsa work (Reynolds, 1982) showed that the concentrations of some of the hydrocarbon species at upper levels (above 500 m) were less than half their concentrations at ground level. A relative profile was used to approximate the decrease in hydrocarbons aloft at the top boundary of the model domain. Ozone predictions at the monitoring sites were reduced by around 10 percent on day 79180 as seen in Figure 30, except at Highland in the vicinity of the predicted peak where the reduction was greater. The daily maximum predicted ozone was reduced by 16 percent.

At a later date, the sensitivity of the model to incoming hydrocarbon concentrations was investigated using surface data from the Pawnee Grasslands summer experiment run by NCAR (Delany, 1981; Greenberg and Zimmerman, 1982). Two situations were investigated, one representing clean air background (.02 ppmC NMHC) and the other representing "dirty" air (1.2 ppmC NMHC at the upwind boundary). The original simulations used an intermediate background (.05 ppmC NMHC). Two different days were simulated, one with the daily peak ozone cloud located beyond the urban area (79180) and one with the daily peak ozone cloud located over the urban area (80204). When the ozone cloud was outside of the urban area, day 79180, the daily ozone maximum was reduced by 12% in the clean air sensitivity and increased by 15% in the dirty air sensitivity. When the ozone peak was over the urban area, day 80204, the change in the daily maximum was -1% and +1% for the clean air and dirty air sensitivities, respectively. Thus, there was not much influence when the ozone peak occurred over the urban area.

The suspected effect of possible interpolation-related errors in the wind fields and possible errors in the area source emissions inventory deserve mention. A trajectory analysis of the peak predicted by EPA2 for day 80204 suggested that there was a narrowly confined slow-down and 180° reversal in the wind field (a "dead-spot") for several hours that influenced the cell containing the predicted peak. The resultant ozone peak is high and narrow, yet the monitoring data suggests a large, broad peak. The wind trajectories of the surrounding cells do not show this "dead-spot" behavior. As well, EPA1, with its "leaky" horizontal diffusion does not show this behavior of predicting a high, narrow peak as does EPA2 (see Figure 23 and the section on dispersion). EPA1 predicts a broad peak more in accordance with that suggested by the monitoring data. More analysis than was possible in this study would be necessary to judge whether this section of the wind field is simply improbable in its behavior or is an artifact of interpolating the wind fields. In any case, it is clear that EPA2 is much more sensitive to such errors in the windfield data set than EPA1.

Isopleths of the ozone predictions consistently show a "hole" or narrowly confined low point in the predicted ozone just below CARIH. An example was shown in Figure 16. Contour plots of non-methane hydrocarbon and NO_x emissions (Figure 24) show a large source of NO_x in a single adjacent cell. Thus it is expected that this NO_x source is affecting the magnitude and spatial extent of the predicted ozone peaks. This spot of NO_x in the area source inventory may be a major contributor to the bias at CARIH. Sensitivity studies are needed to determine the magnitude and extent of the NO_x source's impact on the predicted ozone.

Dispersion (Vertical Mixing Rate) Influence

While the vertical mixing rate is not a variable that is easily accessible to the modeler, it was known from previous carbon monoxide modeling with the equivalent of the EPA1 model (see Figure 31) that the model had a serious problem of underprediction when unstable conditions were being modeled. It seems that the Lamb polynomial calculates a diffusivity near the surface that is almost an order of magnitude too large for even free convection situations (Panofsky, 1981). McRae (1981), in his thesis, took similar note of the problem and changed the Lamb polynomial near the surface. A discussion of this is included in the Masters thesis of Robbi Keil (1983).

The daily bias in carbon monoxide predictions can be useful to isolate dispersion effects in the model because, unlike ozone, carbon monoxide concentrations are not confounded by chemical reaction effects. The bias in the total non-methane hydrocarbon (NMHC) predictions are also useful for this purpose. While NMHC is certainly not inert, ambient levels are dominated by relatively less reactive paraffinic compounds.

Figure 32 shows hourly bias, and Figure 33 shows daily bias versus observed concentration, indicating that, indeed, there is underprediction of CO and NMHC. Thus a bias exists that could be associated with too rapid a vertical mixing rate, using CO and non-methane hydrocarbons (NMHC) as the indicators. The shape of the all-site hourly bias for CO in Figure 32 is different than the shape of the all-site hourly bias for ozone of Figure 9. The ozone hourly bias has an additional hump at mid-day, suggesting other factors also contribute at this time.

Given that CO is relatively inert, the observed underprediction can either be explained by an inventory problem or by the pollutants being mixed too rapidly upward, away from the surface. If there were an inventory problem, one would expect the hourly bias to be worst at the time of the CO maxima, that is, during the morning rush hour (hours 7 and 8 in Figure 32). This is not the case. The absence of significant bias until mid-morning in the all-site average for CO suggests that too rapid a vertical mixing is a more plausible explanation. Examination of the algorithm used to calculate vertical diffusivity in the model corroborates this explanation.

While no action was taken for this work, a sensitivity study on two days was undertaken to investigate the impact of reducing vertical mixing on the ozone predictions. The vertical mixing was reduced from the mixing occurring for a normal day with free convection by assuming a neutral atmosphere for the entire day. The two days picked were a day with the ozone cloud on the outskirts near Highland (79180), and a day with a broad ozone cloud in the central part of the urban region (80204).

The main illustrative results for ozone are reproduced in Table 17. They show that decreasing the vertical mixing increases the unconstrained (full grid) predicted daily maximum for ozone. It also increases the site maxima when an ozone cloud is at (or nearly at) the site. The hourly difference, C_O-C_P, for a monitoring station at the time when there is an ozone maximum is considerably improved. The daily bias, on the other hand, is hardly changed. Opposing changes occur in the hourly biases that are averaged out in the daily bias.

The time series of the predictions for the regularly simulated day and the neutral day are shown in Figures 34 to 37 for CO and O_3 both. EPA1 was used for this sensitivity analysis because that model had been used for the CO modeling. As can be seen, reducing the vertical mixing does make a difference, especially in the region of the ozone cloud. For Day 79180 in which there is no ozone cloud at the center, there is very little difference in the prediction for the central monitoring stations, but there is a large difference for Highland, near the ozone cloud.

Greater detail is given by looking at the hourly differences for all stations for the two days. These are given in Table 18. As can be seen, the hourly differences increase somewhat in the late morning and late afternoon when going to the neutral day. The differences for the hours of the peak improve for those stations near the peak. When all stations are averaged, Day 79180 shows little change in the hourly bias, but Day 80204 shows good improvement at the peak hours. The daily bias is very little changed when comparing the regularly simulated day with the neutral day. Decreased vertical mixing results in less ozone being entrained from aloft, leading to slightly lowered ozone predictions at stations away from the peak. This counterbalances the greater ozone production from emissions in the vicinity of the peak, resulting in little change in the daily bias statistics.

The results for carbon monoxide show considerable improvement. The daily bias of CO on day 79180 is reduced from an underprediction of 42% of the observed concentration to 30%. For day 80204 the bias changes from a 13% underprediction to a 27% overprediction. The time of most interest is mid-day, from 1000 to 1500, when the mixing is strongest. The mid-day bias of CO on day 79180 was reduced from an underprediction of 71% to one of 52% and on day 80204 reduced from an underprediction of 48% to just 2%. From this sensitivity study, we conclude that some portion of the extreme bias at mid-day for CARIH, Arvada and CAMP very likely can be attributed to too rapid a vertical mixing of the reacting pollutants away from the surface. The results presented here suggest that excessive vertical mixing contributes to the bias found in the model predictions. As discussed later, vertical mixing also has an important influence on the predictions of peak ozone when there is a change in emissions.

Missing the Peak in Time

Two examples of the models predicting the local peak at the wrong time are shown in Figure 38. In the 11-day sample, the local site maximum was predicted at the wrong hour by all three models more than 60% of the time.

This mis-timing of the predicted peak has an effect on the daily bias and noise which may be confusing. If the magnitudes of the predicted and observed peaks are similar, predicting the peak too late as in Figure 38 will produce a positive difference at the hour of the observed peak and a negative difference at the hour of the predicted peak. These will tend to balance out in the daily bias computation, producing low bias but high noise.

Examination of the CAMP data for each day shows mis-timing of the peak to be a common occurrence at that site. It explains the pattern of hourly bias at CAMP shown in Figure 7, where noontime bias is positive and afternoon bias is negative.

The pattern in Figure 38 for CARIH is actually atypical, however. Day 79218 is the only one in which the magnitude of the predicted peak approaches that of the observed peak. On other days, the predicted peak is much too low (as well as usually occurring at the wrong hour). Therefore there are no negative differences to balance the positive ones in the daily bias computation.

E. COMPARISON OF DAILY MAXIMUM CONCENTRATIONS

Model behavior was probed in detail above for diagnostic purposes to search for a variety of causes of error. Comparison of observed and predicted concentrations over all of the daytime hours were necessary for a complete diagnosis. However, this very detail tends to obscure the questions which are of most importance in a regulatory evaluation. There a primary concern is with the daily maximum concentrations; therefore, we must evaluate the models' success in predicting the daily maxima. We will look first at the local site maxima, then at the overall daily maxima.

Local Site Maximum for Each Day and Site, Paired by Hour

The most stringent pairing of daily maximum concentrations matches the observed maximum at each site for each day with the predicted concentration for that hour at that site. Five monitoring sites and 11 days result in 55 paired observations. Frequently the maximum prediction misses the time of the observed maximum by one to three hours. Therefore underprediction is to be expected with this pairing, even if errors are merely random.

The performance measures for our three models under this pairing method are shown in the upper half of Table 19. For all three models, the bias estimate is approximately 40% of the mean observed site maximum. That is, the models tend to underpredict by about 40%. Noise levels are similar for the three models as well. Furthermore, predictions of all three models have significantly smaller variances than the set of observed daily maxima. Although a t-test at the 95% confidence level does not show a significant difference in bias between the three models, the nonparametric

Wilcoxon test shows the difference in bias between the DOT and EPA1 models to be statistically significant, with z = 3.21.

Local Site Maximum for Each Day and Site, Unpaired by Hour

Removing the requirement that the model predict a day's maximum at a given site at the correct hour, this pairing matches the observed maximum with the maximum prediction over all hours for that day at that site. Again, there are 55 paired observations, each day's maximum at each site.

The lower half of Table 19 shows the performance measures for this pairing method. The bias estimates have decreased to approximately 30% of the observed mean; that is, the models underpredict by about 30%. For each model, the change in bias from the first, more stringent, pairing is statistically significant ($\alpha = .05$) using the t-test. The t-test again does not show a significant difference in bias between the three models, but the Wilcoxon test shows both EPA1 and EPA2 to have significantly smaller bias than the DOT model (z = 3.43 and 3.21, respectively). Evidently, even with a sample size over 50 the nonparametric test is the more sensitive one when data differ greatly from a normal distribution.

Noise levels are quite similar for the three models, and are similar to those obtained in the first pairing. Variances of the predictions remain significantly smaller than variances of the observed site maxima.

Separate analyses for each site were performed using this pairing method. Arvada and Welby sites produced results similar to those for the data set as a whole, with bias estimates near 30% of the observed mean for all three models. Predictions at the CARIH site are more biased, with average underpredictions ranging from 35% to 43% in the three models. At CAMP and Highland, estimated biases are lower, ranging from 14% to 27%. Most biases are significantly greater than zero under both Wilcoxon and t-tests, the only exceptions being the EPA2 model at CAMP and all three models at Highland.

Noise levels are slightly lower at CAMP, CARIH, and Welby than in the data set as a whole, and considerably higher at Highland. Noise differences between models are small, but differences between sites are more pronounced, with Highland predictions having significantly higher noise than several other sites. (Highland's high noise was the result of missing the location of the peak on the three days when the peak occurred at Highland, as discussed above in the section on hourly predictions.)

Another distinction between sites is in the variability of the predictions. At Highland, predictions of all 3 models have significantly smaller variances than the observed data (probably explained by missing the peak in space, as described earlier), while at CAMP, prediction variances are not significantly different from observation variances for any model. At the other three sites the models perform differently: the DOT model predictions are significantly less variable than the observed data, while the EPA2 predictions are not. In fact, at CAMP, CARIH, and Welby the EPA2 model achieves variances extremely close to the observation variances. Unfortunately, high bias and noise levels indicate that this variability is frequently occurring on the wrong days.

The low variability in the model predictions can be attributed to a number of factors which have already been discussed in regard to introduced errors. Many input parameters were held relatively constant from day to day, including background concentrations, photolysis rates, and emissions. A priori, these would be expected to lead to lower variability in the predictions. Given that these factors were common to all three models, other factors clearly must be contributing to the lower variability in the DOT and EPA1 predictions as compared to EPA2.

Correlations between observed and predicted site maxima range from −.122 to .622 over the 5 sites and the 3 models. Only one of the correlations is significantly different from zero, however, (EPA2 model at Arvada) because of the small sample size at each site (critical value is $r_{.05} = .576$ with d.f. = 10).

The above two sets of comparisons for maxima on each day at each site have examined the question, "how well are region-wide ozone maxima for the day reproduced by the model for the area covered by the monitoring stations?" This question is part of the evaluation of the model's ability to replicate the bulk production of ozone which is of central regulatory concern. The focus of regulatory concern, however, is centered on the bulk ozone production at the major peaks, since only the peak prediction is used in SIP analyses. Thus we next examine the daily maxima in terms of each day's peak concentration.

All-Station Daily Maximum, Paired by Site

In this comparison, the observed maximum from any monitoring station for a given day was paired with the maximum predicted on that day at that site (Comparison (a)). There was no pairing by hour. If there are errors in the spatial positioning of the prediction, then constraining the predicted maximum to a single fixed site can be expected to lead to underprediction by the models by chance alone.

Statistical comparison of the three models on this data is shown at the top of Table 20. The estimated bias ranges from 43% to 49% of the mean observed maximum, $\overline{C_o^{max}}$, for the three models. A t-test does not detect a significant difference in bias between the three models. The Wilcoxon test, however, indicates with 95% confidence that the bias in the DOT model is significantly greater than that in EPA1 and EPA2 on this data. Using the Friedman test to compare the three models jointly also indicates a significant difference between models (chi-square = 6.682, d.f. = 2, p = .035). Results of the Wilcoxon test on this data are shown in the upper part of Table 21.

Variances of the predictions are considerably smaller than the observation variance for all three models (although only for the DOT and EPA1 models are the differences statistically significant). A scattergram of the all-station daily maxima, paired by site, is shown in Figure 39(a).

These features of poor performance (high bias and low variability) are tempered somewhat in the EPA1 and EPA2 models by relatively low noise and high correlations between C_o and C_p.

All-Station Daily Maximum, Unpaired by Site

Removing the requirement that the model predict the daily maximum at the correct site, the observed maximum is paired with the predicted maximum for that day at any monitoring site (Comparison (b)).

Model performances on this data are summarized in the middle of Table 20. The estimated bias ranges from 31% to 42% of $\overline{C_o^{max}}$. Again, the t-test does not show a significant difference in bias between the three models. However, the Wilcoxon test at a 95% confidence level indicates that EPA2 has significantly less bias than the other two models. The Friedman test on the three models jointly substantiates this (chi-square = 7.818, d.f. = 2, p = .020). Wilcoxon tests for this data are shown in the center of Table 13. Both EPA1 and EPA2 predictions are positively correlated with the observations, with noise levels somewhat less than in the DOT model. The scattergram of the daily maxima, unpaired by site, is shown in Figure 39(b).

Area-wide Daily Maximum, Over Entire Modeling Region

Observed concentrations are available only at monitoring sites, but predicted concentrations are available at a large number of grid points over the entire Denver metropolitan area. The chance of attaining the true maximum, then, can be expected to be higher over all of the grid points than over the 5 monitoring sites. Thus in pairing the observed maximum with the full-grid predicted maximum, we should expect the model to overpredict (Comparison (c), Table 20). This tendency will be partly counterbalanced, though, by the fact that monitoring sites have been deliberately located in high pollution regions and that our sample consists of days when high ozone concentrations were observed at the monitoring sites.

Statistics at the bottom of Table 20 show that the models continue to underpredict the daily maxima, with biases ranging from 10% to 30%. Here for the first time, however, bias in one model, EPA2, is not significantly different from zero (under both Wilcoxon and t-tests). In comparing models, the t-test indicates only that bias in the EPA2 model is significantly lower than in the DOT model at the 95% confidence level ($t = 2.57$). The Wilcoxon test detects distinctions between all three models, with EPA1 significantly less biased than DOT, and EPA2 significantly less biased than either of the others as shown in the lower part of Table 21. Comparing the three models jointly on this data, the Friedman test shows highly significant differences (chi-square = 14.045, d.f. = 2, $p = .001$).

Variability in the EPA2 predictions is very close to the variability in the observations, with a significant positive correlation between the observed and predicted maxima. The scattergram of area-wide maxima, unpaired in space, is shown in Figure 39(c).

Regression Analysis of the Daily Maximum Pairings

A linear regression was performed on the three different pairings shown in Figure 39 (paired by site, unpaired by site, and unconstrained in space). The results, giving the slope, the intercept, and the coefficient of determination (r^2), are shown in Table 22.

Tests of significance of the correlation and regression coefficients require that the data be normally distributed. Because this data is probably not normally distributed, we have used the coefficients only as rough indicators rather than as statistical tests. The significance test dramatizes the imprecision of a correlation computed on only eleven points, however. The sample $\hat{r}$ must be above .576 (or r^2 above .332) in order to be significantly greater than zero. Thus small differences in correlation between two models should not be given undue importance. Table 22 shows that the DOT model has a consistently lower correlation between observed and predicted peaks than the two EPA models. The differences between EPA1 and EPA2 correlations are relatively small. The slopes and intercepts in Table 22 do not provide a way to select between the three models. In this example, the scattergrams in Figure 39 provide a better view of the differences between the models and their relationship to the line $C_o = C_p$.

Evaluation of the Models Based on Daily Maxima

Substantial underprediction of peak concentrations has been shown for all three models. A general impression emerges that the DOT model is much less able than EPA2 to predict day-to-day changes in peak ozone concentrations, while EPA1 falls somewhere between.

In every comparison, the DOT model shows the highest bias of the three models, and a near-zero correlation between observed and predicted maxima. In addition, the variability of DOT predictions is extremely low: significantly lower than the variability in the observed maxima in every comparison. Despite this low variability, the "noise," S_d, in DOT predictions is slightly higher than that in the other two models. Thus the variability it does produce is more likely to occur at the wrong times and places. We conclude that the DOT model is less capable of predicting peak ozone concentrations than either the EPA1 or the EPA2 model.

Differences between the EPA1 and EPA2 models are smaller. Correlations between observed and predicted maxima are similar for the two models. Using the all-station, unpaired daily maximum, (comparison (b), Table 20), the values of r^2 indicate that EPA2 explains 46% of the observed variance while EPA1 explains 36%. Bias in EPA2 is somewhat smaller than that in EPA1 in most cases, while variability in EPA2 predictions is consistently higher than in EPA1 and closer to the variability in the observations. Many of these differences are not statistically significant, but the consistent slight superiority of EPA2 on all of the performance measures indicates that it performed best of the three models on this data set.

Whether the performance of the best model is "good enough" for regulatory purposes is still a question requiring professional judgment. The substantial bias in the EPA2 model can reasonably be estimated to be between 10% and 31% of observed maxima depending on whether predictions are confined to monitoring sites or selected from the full grid. This could indicate a need for some kind of calibration or tuning of the model, unless improvements in the input to the model are found to correct the bias. Using predictions from all grid points (unconstrained in space) appears to be a way to reduce the bias in the model. This might be appropriate if the underprediction has been caused by errors in location of the peak due to errors in the wind field.

Similar results were obtained in a study of the performance of the EPA2 model on ozone concentrations in the St. Louis, MO, area (Cole, 1982a). That study concluded that all grid points should be used in determining the predicted maximum. When this was done, the researchers found that predicted peaks for most days were within ±30% of the observed peaks and concluded that the model performed with a "reasonable degree of accuracy" in estimating observed peak ozone. By this judgment, the EPA2 model performed reasonably well on the Denver data too, since predictions for all of our 11 days fell within ±27% of the observed peaks.

F. EMISSIONS CHANGE COMPARISON

The attempt to gain a better understanding of the performance of the model by diagnosing errors showed that a complex of information is contained in the bias and the other measures of paired comparisons. The analysis of errors led to a multitude of answered, partially answered and unanswered questions about how well the model is performing. The complexity of the evaluation and list of probable errors left an incomplete appreciation of the strong and the weak points of the model, not enough for an unambiguous assessment, to our minds, of the acceptability of the model for regulatory purposes.

Then the comparisons of daily maxima, while addressing more directly the manner in which the model would be used for regulatory purposes, raised new doubts about the performance of the model. These doubts, raised by such results as the low variability in the model predictions and the unsatisfactory correlation and regression coefficients, created a serious question as to whether it is valid to draw inferences about the performance of the model under conditions of changing emissions, based on a performance evaluation in which the emissions do not change. In the same vein the sensitivity studies for the evaluation and for the data set development showed how complex the model predictions are and how non-linear the changes can be.

The ultimate goal of a performance evaluation such as this one is to answer the question, is the Urban Airshed Model good enough for regulatory application? Can a bias in the predictions for a set of historical days be calibrated out with any confidence that the predicted changes due to changes in emissions are still valid enough to be used? We believe the performance evaluation as carried out thus far is incapable of giving a sufficiently unambiguous answer to that question. That question must, we believe, be tested directly for photochemical models. In this section, therefore, we present one approach for directly testing the predictions of the photochemical models in order to evaluate their response to emissions changes.

The approach which we have taken is to assemble a second, complete emissions inventory for an earlier year. That earlier year had to be sufficiently separated in time from the year of the performance evaluation data set so that changes in the ambient concentrations, associated with changes in emissions, had actually been observed. The meteorological conditions of the set of performance evaluation days was then used to re-simulate a set of pseudo-days using this "new" emissions inventory. This procedure predicted changes in ozone concentrations due only to emissions changes for a fixed set of meteorological conditions. The relative change in the predicted pollutant concentrations was then compared with the relative change determined for the observed pollutant concentrations.

For photochemical models, because the chemistry and meteorology are highly interrelated, and because the performance of the models could be different for different ratios in the emissions of NO_x to non-methane hydrocarbons, the ideal test approach would be to also perform a symmetrical evaluation to the one just described. That is, a second set of evaluation days should be established for the earlier year. Then an emissions change test would be performed again using the later emissions inventory to re-simulate pseudo-days associated with the meteorological conditions of the evaluation days of the earlier year. Thus the analysis presented here is one-half of a more ideal analysis approach for testing a photochemical model. It should, however, represent an adequate approach to testing the predictions of models. In either case (i.e., each half of the ideal evaluation), the use of pseudo-days is necessary because exact replicas of meteorological conditions in two different years would be nearly impossible to find. The approach also replicates the conditions under which the model will be used in regulatory analysis.

Definition of an Emissions Change Comparison

The emissions change comparison should resemble as closely as possible the manner in which the air quality models will be used in a regulatory application. Thus we are interested in the changes in the area-wide (unconstrained) and all-site daily maximum predictions that occur due to a change in the emissions. For the model simulations, that change in ozone concentrations is represented by use of the two emissions inventories, with meteorological conditions held constant. For the monitoring data, it is necessary to find a way of separating the effect of emissions changes on the ozone trend from the effect of differences in meteorology from year to year.

Estimating these changes is not a trivial task. A change in emissions implied by the difference between two emissions inventories is only valid if the techniques used to estimate the emissions in each inventory are the same. Otherwise changes in techniques must be corrected for. Trends in ambient concentrations can be confounded or even masked by such things as changes in the location of a monitoring station, changes in the chemical technique used to measure concentrations, changes in calibration procedures and, last but not least, year-to-year meteorological variability. All of these factors must be checked for and taken into account in any trend analysis of ambient concentrations.

Development of the Earlier Emissions Inventory

Because the increase in the vehicle miles traveled (VMT) in the Denver area has been so rapid between 1970 and 1980 (4.7% per year), it takes several years for the reduction in automobile tail-pipe emissions to have a noticeable effect on total Denver emissions. Thus the two emissions inventories should be several years apart. Availability of a transportation data base can be a severe limitation on the choice of years, however. The earliest year for which a transportation data set was available for Denver was 1975. The earliest year of the most reliable transportation data set, i.e., the transportation data set which had the most up-to-date corrections, was 1976, because that year was the base year of the 1982 State Implementation Plan (SIP) projections for ozone. As well, a point source inventory had been developed for 1976 as part of the SIP work. The mobile source emissions model of EPA (MOBILE2) was expected to be reliable for any of the earlier years. Thus 1976 was chosen as the year for which to develop the second emissions inventory.

Because of the rapid growth in Denver's VMT, the span of 1976 to 1979 was considered to be barely long enough for this test and could turn out to be marginal. It was the best that could be achieved, however. An emissions inventory representing 1976 traffic and emissions conditions was developed for both the Carbon Bond I and Carbon Bond II hydrocarbon splits.

The 1976 inventory was very comparable to the 1979 emissions inventory. Both inventories used the same large-scale transportation model to establish the location and magnitude of the vehicle miles traveled. Major traffic count programs had been carried out in 1975 and 1979 in Denver to help adjust the transportation model results. Both inventories used the same mobile source emissions model and the same procedures to estimate the automotive emissions for given vehicle miles traveled. The 1979 point source inventory was part of a periodic update of the 1976 point source inventory. Thus problems or bias that might be associated with the emissions inventory would be systematically similar for each inventory.

Choice of Models and Days for the Emissions Comparison

The major purpose of the emissions comparison presented here should be a demonstration of its importance and usefulness. The question that ought to be answered is, does the emissions comparison provide us with new information that we did not already have in some form from the above hourly and peak-concentration comparisons? Therefore, it is important to perform the emissions comparison on all three of the air quality models.

Ideally all 11 performance days should have been re-simulated. Because computer resources were limited, however, the number of days re-simulated with the second emissions inventory had to be reduced to eight. The choice of the three days to exclude was based on EPA2's performance for the unconstrained daily maxima. We elected to pick days with reasonably consistent gross error performance. We also wanted them to span the range of the observed maxima. Of the daily maxima comparisons Day 5 (79249) had the greatest underprediction (26%) and Day 10 (80207) had the greatest overprediction (27%). In fact, Day 10 was a day in which the major ozone peak was most likely not observed at any of the monitoring stations. Concomitantly, these days also had the largest percent absolute deviation (see Figure 4). Thus they were considered to be less typical of the average gross error performance of the model. Removing these days would not affect the range of our predictions, thus they were excluded from the test.

The third day to be excluded was Day 7 (80177). It was a toss-up between Day 7 and Day 6 (80170). Both had the same observed daily maximum and nearly the same predicted daily maximum. Eliminating one should not cause much of a loss of information. As shown in Figure 3, Day 7 had the larger percent absolute deviation of the two days and it was slightly less characteristic of the average gross error performance of the model, thus it was excluded.

Estimation of the Change in Observed Maxima

Ideally, a regression or time-series analysis of several years of data that is based on a stochastic model with an accurate deterministic component should be used to most precisely estimate the underlying trend in the data. The trend due to emissions changes needs to be separated from the year-to-year variation in the meteorology. Work of such a nature on the Denver data set, independent of this research effort, was not advanced enough to use at this time, nor were we aware of other available work on this problem. Thus simpler techniques to establish the trend had to be used to carry out the evaluation of the emissions change comparison.

The earliest year for which ozone data at the five monitoring stations exists is 1975. Thus the trend had to be established using data for the 1975 through 1980 time period. The distribution of the ozone monitoring data of Denver is cube-root normal. Thus one could not assume that an annual trend in ozone concentrations derived from monthly means would be the same as a trend based on just the extreme end of the distribution of ozone concentrations. The trend in the observed daily maxima needed to be computed using a subset of each year's observed daily maximum concentrations which resembled the limited population of days in the performance evaluation data set. The fraction of the concentration distribution used needed to remain constant from year-to-year in order to give the same weight to each year's data. Therefore, the range of concentrations defined by the concentration cutoff of 100 ppb had to be replaced with an equivalent definition for the range in terms of a percentile of the distribution of daily maxima. The same percentile of the daily maxima for each year needed to be used in the trend analysis, not the same range. Thus, for each year, the same number of daily maxima is used in establishing the data set of observed high ozone concentrations.

Two different high ozone data series were established for the trend comparison as a sensitivity check. The first data series was the top 11 days of each summer's observed daily maxima, with consecutive high ozone days excluded to better resemble the evaluation data set. Eleven days were used for this set to make the number comparable to the number of evaluation days. For each year, the eleven days having the highest peak concentrations were used, after excluding all but the first day of any multiple day series of high ozone. The second series was the top 14 days of each summer's daily observed maxima, with consecutive high ozone days included. Only 14 days from each year were used because there were only 14 daily maxima of at least 100 ppb in the summer of 1980. The annual means of the two high ozone data sets are shown in Figure 40.

The large scatter from year-to-year is evident. The scatter is not due to any changes in monitoring location or chemical techniques. A national change in calibration techniques took place in 1978, affecting the 1979 and 1980 data. For Denver, that change was found to be minor, a slight increase in ozone reported of at most 3.8 percent. There was apparently no bias before or after the calibration change. The change just reduced the random error. It was not possible to draw any simple association between the scatter in the high ozone readings and the year-to-year variability of the wind speed. Thus external information had to be used to infer a best estimate of the shape of the annual ozone trend before any fit could be tried through the points in Figure 40.

To determine the appropriate shape of the ozone trend, the trend in daily hydrocarbon and NO_x emissions was investigated. The trend in hydrocarbon and NO_x emissions was based on the point source inventory trend established by the Colorado Department of Health, the average daily VMT estimated by the Department of Highways for each year (based on traffic count data) and the mobile source emission factors for each year from EPA's MOBILE2. The resulting daily NO_x emissions were expected to increase somewhat over the period 1975-1980. The resulting daily HC emissions showed a near-perfect straight line decrease from 1975 through 1979. 1980 had double the decrease because VMT did not increase that year, due to high gasoline prices. The 1980 VMT was 5 percent lower than would have been expected, assuming a regular, smooth trend from 1975 to 1980. An examination of an EKMA isopleth indicated that the change in ozone expected per unit decrease in HC would lessen slightly as the hydrocarbon emissions decreased. The conclusion from integrating the above information was that, while it would not be perfect, linear regression of the data points forming the averages shown in Figure 40 was considered to offer a reasonable approximation of the trend in ozone concentrations that could be attributed to decreasing emissions with time. The linear trend analysis on the top 14 daily maximum ozone concentrations for each year, 1975-1980, shows that they have decreased at the average rate of 6.2 ppb per year. This rate of decrease is significantly different from zero at the 99% confidence level. The trend analysis results are virtually identical for the top 11 days of non-consecutive ozone daily maxima for each year.

Results

The predicted percent increase in each day's peak ozone due to an increase in emissions, corresponding to the change from 1979 to 1976 emissions conditions, is given in Table 23 for each model. Clearly the three models seem to show a different response to changes in emissions. Because the performance evaluation emissions inventory corresponds most closely to 1979 emissions conditions and the second emissions inventory to 1976 emissions conditions, the percent increase in observed ozone maxima should be computed using a 3 year interval. Table 24 shows the estimated trend in the observed ozone concentrations, as well as the trends due to emissions change that are predicted by the three models. The mean of the observed concentrations on the 8 days used in the emissions change comparison was 137 ppb. Under the same weather conditions, then, on the basis of the trend in observed concentrations we would expect the mean concentration to have been 18.6 ppb, or 13.6% higher in 1976.

For each model Table 24 gives the predicted rate of change in the daily maxima, both when the maximum is selected only at monitoring sites and when the maximum is selected from the full grid. To allow for the bias in the models, these changes should be compared as a percentage of predicted concentrations, rather than in absolute form. For each model, the average change in peak ozone predicted between 1976 and 1979, expressed as a percentage of the mean predicted peak ozone concentration for the 8 sample days, is shown in the last column of Table 24. No matter which way the predicted daily maxima are selected, the EPA2 model is more responsive to the emissions change than either the EPA1 or the DOT model. Even EPA2, however, does not predict as great a change as that found in the observed ozone data.

None of the differences between the models in Table 24 are statistically significant. Still, on the basis of this comparison and the earlier comparison of the models on daily maxima, we would conclude that EPA2 is superior to DOT and EPA1 for regulatory purposes.

All models performed better than we had expected, based on the performance evaluation up to this point. The implication is that the results of earlier segments of the evaluation did not give a good indication of how the models would perform under a change in emissions. The emissions change comparison produced and highlighted new, independent information. The DOT model daily maxima predictions showed no effective response to the differences in meteorology (the correlations and regression coefficients in Table 20). Yet, the DOT daily maxima did show a response to changes in emissions.

None of the performance comparisons before the emissions change comparisons gave any indication that DOT would perform two-thirds as well as EPA2 on the crucial test for regulatory purposes. The hourly comparisons at the monitoring sites indicated there was little significant difference between models. The peak-concentration comparisons indicated there was some difference between models (see Tables 20 and 22). As illustrated in Figure 41, there is no relation between daily bias and the percent change in ozone predictions due to a change in emissions. As the scattergrams of Figure 42 indicate, there is little or no relation between either the observed or predicted peak ozone concentration for a day and that day's percent change in the daily maximum ozone due to the emissions change. Although Figure 41 shows no relation between daily bias and sensitivity to changes in emissions, daily bias has been shown previously to be a rather insensitive measure of effect. Looking instead at the bias in the peak predictions (Figure 43), there does appear to be some relationship between bias and emissions sensitivity, at least for the EPA2 model. Days with the highest bias show a smaller percent change in the peak ozone predictions. This is consistent with the slight underestimate of the slope of the trend line in the ozone observations over the 1975-80 time period. Results from the vertical mixing sensitivity tests to be discussed later suggest a possible mechanism for this effect.

A rather important piece of information is evident in the results just presented. The modification to the Urban Airshed Model that produced the greatest improvement in its predictions on the emissions change test was not the modification that produced the greatest improvements in its predictions for changes in meteorology. This can best be seen in the comparison of peak predictions for the unconstrained pairing. That pairing is least sensitive to the distortion that is introduced when the monitoring stations happen to be consistently outside of the predicted cloud of high-ozone. Table 24 implies that the elimination of the horizontal numerical diffusion (the modification from EPA1 to EPA2) produced the great majority of the improvement in the predictive capability of the model for a change in emissions. Tables 20 and 22 and Figure 39 imply that the change of chemical mechanism (the modification from DOT to EPA1) produced the majority of the improvement in the predictive capability of the model in response to a change in meteorology with a "fixed" level of emissions.

The above results suggested the need for a sensitivity study on the effect of vertical diffusion on predictions of the model when emissions change. It was observed above that reduction of the rate of horizontal diffusion (the modification from EPA1 to EPA2) increased the predicted relative change in peak ozone for a given change in emissions. This raised the question, "would a change in the vertical rate of diffusion affect the predictions similarly?" The Urban Airshed Model does appear to have a problem with too rapid a vertical mixing. Resource limitations precluded a full sensitivity study, but two days that were included in the emissions change comparison, 79180 and 80204, had also already been simulated with a reduced vertical diffusivity using EPA1. Therefore those two days were resimulated using EPA1 with the 1976 emissions inventory and reduced vertical mixing. This allowed the calculation of a new trend in peak ozone resulting from changing emissions, for the case of reduced vertical mixing.

Reducing the vertical diffusivity in EPA1 increased substantially the predicted relative change in the daily ozone maximum for the given change in emissions. The relative response of the ozone peaks to changes in emissions increased similarly for the two days with reduced vertical mixing, a somewhat different response than for the change from EPA1 to EPA2. With reduced vertical diffusion, the relative change in ozone peaks between the 1979 and 1976 data sets increased from 5.9% to 10.8% and from 10.5% to 19.9% for days 79180 and 80204, respectively, which are substantial and nearly identical percentage increases in the predicted change for the two days. This looks like a very important effect and should be directly checked with EPA2's predictions, since both the horizontal and vertical rates of diffusion at the surface seem to be important.

The error in advection for day 80204, which probably caused an abnormal peak in EPA2, did not seem to affect the prediction of a change in ozone due to a change in emissions. The two EPA model versions give fairly similar predictions for day 80204, 10.5% and 14.0% for EPA1 and EPA2, respectively. In addition, a 14.0% change is a typical prediction for EPA2 on the eight days. Thus it appears that although such errors in the wind field will introduce a bias in the model's prediction for a given year, that effect is not necessarily carried over into the relative predictions of the model for an emissions change. This is an area that deserves more investigation.

The above analysis leads to two basic conclusions. First, if the intent of the performance evaluation is to assess the acceptability of the air quality model's projections of the effect of emissions changes, then that assessment has to be made directly by testing the model on a data set which involves a change in emissions. Inferences regarding model performance with respect to emissions changes, if based on data sets which do not involve a change in emissions, will be unreliable and potentially misleading. In addition, inferences about areas on which to focus model improvements will also be misleading. For example, the single year's evaluation with respect to meteorological change seemed to suggest that future effort should be on further improvements in the chemistry, while the multi-year evaluation with respect to an emissions change seemed to suggest that future effort should be on improving the correctness of the diffusion algorithm in the model. These are two quite different components of the model. This conclusion merely reconfirms that the design of the evaluation has to match the purpose of the evaluation.

Second, it appears that some errors that contribute to the bias in the model's prediction of ozone within a given year will also affect the model's prediction of the effect of an emissions change. It would appear that there are other errors, however, that do not affect the prediction of response to emissions change. More work should be done to understand which errors impact the predictions related to a change in emissions and which do not. This will affect the choice of simulation days and guide regulatory use of the model.

There are two specific areas for further work which are highlighted by the Denver emissions change results: (1) vertical mixing and (2) downwind location of the peak relative to the major emissions sources. In regard to vertical mixing, the rate of diffusion out of the ground-level boxes (both horizontally and vertically) has been shown to greatly influence the predictions of a relative change in peak ozone due to a change in emissions. The sensitivity of the peak ozone predictions to the rate of diffusion is greater for the emissions change prediction than for the meteorological change (single year) prediction. The neutral day used in the dispersion sensitivity study helped, but did not completely remove the bias on days 79180 and 80204. However, the neutral day appeared to overcorrect the emissions change predictions (obtaining a 20% relative change on day 80204). Thus only looking at ozone bias may not be the best way to "get the diffusion right." The fact that the daily CO bias on day 80204 went from underprediction to overprediction suggests that the inert pollutants such as CO may provide a better key to knowing whether or not the diffusion in the lowest boxes is appropriate or not. For each new city, the model may need to be cross-checked for reasonableness. The use of the CO predictions as one basis for that cross-check, rather than ozone predictions, should be investigated. The first step, however, is to correct Lamb's polynomial which is present in all versions of the model to reduce vertical diffusivity near the surface.

The second area in which further work is suggested is related to the downwind location of the peak. An investigation of the percent change in the main ozone peak due to a change in the emissions showed that the change appeared to be affected by the time and location of the peak. The percent change tended to monotonically decrease after 1200 MST and tended to be less, by a fair amount, when the peak was farther from the center of the urban area. Across the eight days the average percent change in EPA2's maximum ozone prediction at a given hour due to a change in emissions from 11.8% at 1200 to 10.3% and 8.1% at 1300 and 1400, respectively. An examination of the eight days in Table 23, using EPA2, showed that days 79180, 79193 and 79208 had predicted peaks that were farthest from the center of Denver. As well, 79193 and 79208 were the only days in which the predicted ozone peak was influenced by the point sources. The hours of the predicted peaks were 1300, 1200 and 1200 for days 79180, 79193 and 79208, respectively, not late in time at all. The model seems to predict less change in ozone due to a change in emissions when the peak is later in time and farther away from the main emissions source. To check whether this tendency would hold on another day, the emissions change test was carried out for an extra day using EPA2, day 80207. This day had a predicted peak of 154 ppb, it was at 1400 and it was to the south of Denver, past Highland. The relative change in ozone predicted for 80207 for the daily maximum over the full grid was 8.4%, much like the three days discussed above. This analysis suggests that the model's prediction of a relative change in ozone due to a change in emissions is affected by the timing and location of the peak. This same effect was seen on two days in Tulsa (Layland, 1983), days having predicted peaks close to and far from the city. The veracity of such a prediction clearly must be checked against monitoring data. A first step would be to check the trends in ozone at each monitoring station to see if the stations farther away from the urban center show less of a decline in the ozone trend. A result showing there is no difference in the trend between stations would have important implications for appropriate use of the model and interpretation of its predictions for different types of high-ozone days. While it is possible that transported NO_x emissions and background hydrocarbons could provide a mechanism for this effect, it is also possible that some types of days should simply not be used for simulations for regulatory decision-making.

In summary, we have found that the emissions change comparison does produce an assessment of the Urban Airshed Model that is fairly independent of the meteorological change assessment. Thus for a regulatory assessment, an emissions change comparison must be included to investigate whether the model is good enough for regulatory use. In addition, two areas of concern with respect to use of the model for regulatory decisions have been raised which suggest a need for further investigation.

G. PERFORMANCE EVALUATION CONCLUSIONS

The example performance evaluation has pointed out a number of operating characteristics of the model. These characteristics relate to its general use as an urban photochemical model and to its use in regulatory analysis. Although the Denver example evaluation has a number of flaws, the lessons about the operation of the model are valuable. In particular, it is evident that it is not a simple task to use the model in the support of decision making. We will use the term "the model" to mean the Urban Airshed Model in general and EPA2 in particular, unless otherwise noted.

We do not believe we can make categorical statements in this report about the goodness of the model. The Denver example evaluation has shown that there are a number of errors in the model and in the input data that are explainable and which affected the results presented in the example evaluation. We do believe that the bias shown in this Denver evaluation can certainly be reduced. Thus the Urban Airshed Model is clearly better than the quantified evaluation indicates. A clear, general impression is that the model (EPA2) has come of age. That does not answer the question of whether it is good enough for regulatory application. It has become clear that the model has idiosyncracies and that one evaluation for one city will not answer that question. The purpose of the discussion of this section is to shed light on that question, based on the Denver experience. The emphasis will be on those attributes that relate to use and evaluation of the model for purposes of making predictions for use in decision making.

Performance Character of the Model

Changes to the Urban Airshed Model from the DOT to EPA1 to EPA2 versions resulted in improved predictions. The day-to-day variability of the peaks was improved. The amount of ozone produced at the surface increased, reducing the bias in the predictions of the peaks. The capability of the model to reproduce changes in peak concentrations due to changes in meteorology was greatly improved. The capability of the model to reproduce changes in peak concentrations due to changes in emissions also appeared to be greatly improved. Nonetheless, the model still shows a pattern of chronic underprediction.

The predictions from one model version to the next changed in the regions of the peaks only, rather than everywhere in the grid. The predicted ozone concentrations for the "valleys" and "saddles" between the peaks and for the large flat regions of low ozone remained insensitive to the changes to the model. The size of the base of the ozone peaks remained unchanged. However, because the peaks were higher, the spatial extent of high ozone areas increased as the peaks increased in magnitude. This was true not only for the change between EPA1 and EPA2, but also for the reduced vertical mixing sensitivities.

The location and timing of the peaks were not accurately reproduced and it appears that the spatial extent of the peaks may be underpredicted. There do seem to be a number of factors (wind fields, emissions, and chemical mechanism) influencing the location and timing of the peaks. This study did not attempt to determine the relative contributions of those three factors to this problem of the model. The predicted peak ozone cloud often moves at a different speed than a parcel of air, generally more slowly, indicating that there is a complex interaction going on. Due to the limited number of monitoring sites it is nearly impossible to say anything definitive about the true spatial extent of the ozone peaks. The very sharp peak predicted by EPA2 on 80204 is considered to be primarily an artifact of the wind field and does not represent a problem internal to the model. From the limited monitoring data available, it appears that the steepest observed rate of change of ozone per distance is approximately two-thirds of the average rate of change on the largest predicted peaks. Thus the predicted peaks still seem to be steeper than those observed. As will be discussed below, one contributor to the problem could be that the off-peak production of ozone is not sufficiently large. The inference that the areal extent of the peaks is underpredicted rests on the fact that the evidence for Denver seems to be consistent with the results from the St. Louis and Tulsa studies.

The behavior of the model is very site specific. In the collection of days that were simulated, every site monitored high concentrations of ozone on a few of the days. The hourly comparisons and the sensitivity studies showed that each site had an individualized combination of errors contributing to the bias at that site. Thus it is not straightforward to interpret the bias and noise statistics across monitoring sites and it should not be assumed that comparable errors are contributing to the the statistical measures across sites. For example, the fact that CAMP is located very near a major intersection of downtown arterials may contribute to its low bias, whereas missing the peak in space and basic underprediction of the peaks were the sources of the high bias at Highland.

The predictions of the model are fairly sensitive to some of the input values. The setting of background ozone levels greatly affected the predictions of the model throughout the day. Thus great care must be exercised in setting background ozone levels throughout the day. There is moderate sensitivity of the predicted ozone maxima to background hydrocarbon levels on some days. The sensitivity to special features in the wind fields and emissions is important for interpretation of the model results.

The EPA2 model has become more sensitive to certain errors or features in the wind and emissions input data as the result of elimination of the artificial diffusion. Any "dead-spots" in the wind field will result in a very large prediction of ozone in a single cell if that cell is located in a peak ozone area. The result is an anomalous spike of ozone, distorting the interpretation of the spatial extent of the predicted high ozone region and affecting the estimation of the bias of the model's predictions. In the same vein, the influence of the point sources on the ozone predictions is increased, because the large NO_x concentrations do not disperse as rapidly in EPA2 as they do in EPA1.

The model behaves differently when emissions change than when the meteorology changes. The level of bias estimated for days which have basically the same emissions does not relate closely to the error in the predictions when emissions are changed. The predictions for an emissions change are much more sensitive to dilution effects than predictions for a change in meteorology. The responsiveness of the model to changes in meteorology appears to have no relation to its ability to predict well the relative changes in concentrations due to changes in emissions. This is an important conclusion of this Denver example performance evaluation. It implies that some inferences from past studies which depend on bias measurements to evaluate the performance of the model for regulatory application may not be valid.

The response of the predicted peaks to an emissions change seems to be a function of time and location. This is an area that deserves more investigation. The general pattern of model response during the day seems to be that as the ozone cloud builds to a peak and then slowly declines, the percent change in the hourly prediction due to emissions changes decreases. In what may be a related phenomenon, the farther away the daily peak is away from the major emissions sources, the less change there is in the predicted peak due to a change in emissions. This behavior of the model needs to be verified against observed ozone data to establish whether this is a problem in the model that becomes evident on non-stagnation days.

The model still has room for improvement. Two areas of improvement are immediately indicated by this evaluation. First, the algorithm for the calculation of vertical diffusivity should be corrected to correspond with observations and theory near the surface. This would reduce the rate of vertical mixing in the model, which has been shown to improve the predictions of the model. The model is sensitive to errors in this vertical diffusivity formulation. Second, the box height at the surface should be held fixed. This is necessary to make sure the lower diffusivity is adequately and uniformly taken into account in the calculations. It will also eliminate a source of error due to variation in the volume of the box between hours and variation in the vertical diffusivity calculated for the surface box, when in reality there is no variation. These two changes should improve the predictions of the model both for an emissions change and for changes in meteorology.

Insights on the Regulatory Use of the Model

Evaluating the model for its acceptability for use in decision making requires an understanding of the idiosyncracies of the model--what influences its predictions. Those idiosyncracies must be accomodated or deemed unimportant in order for the model's predictions to be usable and to hold up under scrutiny of a "hostile" audience. The assumption is that the model, because of its complexity and necessary simplification, will continue to be less than perfect. It will probably continue to perform better for some types of high ozone days than for others. While a number of insights about the model have evolved, we focus here on those that are most relevant to use of the model for regulatory purposes and to minimization of the effects of errors and model idiosyncracies on the peak ozone predictions. These insights are associated with simulation of high pollution cases.

The model has difficulty correctly incorporating strong NO_x sources. This characteristic of the model affects the quality of its predictions, both in terms of magnitude and spatial extent of the peaks. It does appear as if this problem affects model predictions that would be used for decision making. Accepted guidelines need to be developed for users of the model, telling them what to do.

The predicted change in ozone peaks seems to be quite sensitive to diffusion out of the ground-level box. Thus simulation of this diffusion needs to be as correct as possible. The sensitivity of the emissions change prediction to vertical mixing is greater than the sensitivity of the peak prediction on the historical day. Thus bias in peak ozone predictions may not be the best guide to assessing whether the diffusion is correct. One procedure to investigate further is the use of the bias in the prediction of inert pollutants, especially carbon monoxide, as a measure of the accuracy of the vertical mixing reproduced by the model. A means of model adjustment based on CO bias may be important to account for urban differences. Such an adjustment might be far more important than any calibration for obtaining correct predictions of the magnitude of the peak for regulatory purposes. The best situation would be that once the vertical diffusivity is correctly simulated by the model no further adjustments would ever have to be made. Further work is obviously required on this topic.

Changes in peak concentrations due to changes in emissions seem to be a function of time of day and also seem to be sensitive to distance of the peak from the major source of emissions. As the distance increases the predicted change decreases. This characteristic behavior of the model must be verified as correct or incorrect, possibly by performing a trend analysis on monitoring data to ascertain if the trend is a function of distance from the main source of emissions. If this distance effect is real, this could have important implications for regulatory decisions. In any case, it has important implications for choice of the days that are most appropriate to be used for regulatory analysis.

There is some indication that the tail of the ozone peak that trails behind may collapse too quickly. This should be investigated. The cause of this behavior may also be affecting the spatial extent of the predicted ozone cloud. It may also be implicated in the apparent decrease in the predicted change of ozone due to an emissions change as the peak is located farther from the main source of emissions. We have not done any analysis to allow us to speculate as to the cause of this behavior, but it seems worth investigating whether there is a connection because of the implications on guidelines for the model's use.

On the basis of this Denver example evaluation and error diagnosis, it appears that some, but not all, of the problems that contribute to the bias in the predictions for a response to change in meteorology also contribute to the bias in the predictions for a response to emissions change. The degree of contribution is different, however. Other problems do not appear to affect the predictions of peak ozone for a change in emissions. This lack of strong association between the two kinds of predictions means that much more attention must be given to evaluating the model in the way it is intended to be used. As discussed above, only using ozone predictions to evaluate the model may be too limiting. Clearly the ozone predictions of the model for an emissions change are more complicated and sensitive than previously imagined and less related to the types of evaluations currently in common use than presently assumed. This would seem to imply that the model is not yet adapted to casual regulatory use. If a single correction of the vertical diffusivity in the model can apply across most urban areas, then it appears possible that most of the remaining bias in the single year's predictions of the model will not seriously affect its predictions for changes in emissions.

IV. IMPLICATIONS FOR PERFORMANCE MEASUREMENT

Conclusions about the performance measures derived from this study are likely to be dependent, in part, on our particular Denver data set and our simulation results. This also is a special evaluation case, in that the three models to be compared represent incremental improvements in one basic model, which aids in interpretation and reduces the amount of analysis required to evaluate the basic model.

A. CONCLUSIONS ON THE USE OF STATISTICAL TECHNIQUES

Evaluation of the Performance Measures

Bias was the most useful measure of the accuracy of the predicted concentrations. In the set of peak predictions it provided a basis for statistically discriminating between models, through use of the Wilcoxon test. When computed on subgroups of the set of hourly predictions it helped to pinpoint systematic errors in time and space. It is useful also in that it offers an ideal standard, zero bias, against which a model can be judged.

We found that proportional bias, i.e., dividing the bias by the observed concentration, was also useful in getting a sense of the model. But it must be interpreted judiciously because the proportional hourly bias looks rather poor early in the morning but doesn't, at low concentrations, really affect the important predictions of the model.

Noise is less interpretable than the bias. Its ideal value, zero, is virtually unattainable because there are practical limits on how accurate a model can be. The same difficulty applies to the gross error measures, $\overline{|d|}$ and MSE. The goal is to obtain a low value, but there is no standard for determining what value is "small enough" for regulatory use of the model. Our three models did not have significantly different noise levels in any of the data sets discussed above. The noise could be useful, however, in selecting between two models which have similar bias: the model with the smaller noise level would be preferred.

Variability comparisons between the observed and predicted concentrations provided useful diagnostic information, despite the fact that they do not require pairing of the observation and prediction. The ideal variability in the predictions would be equal to the variability in the observed concentrations, and this can be tested using the F-test, although confidence levels are only approximate if the data is not normally distributed. In the example evaluation, the models tended to produce too little variability in their predictions, probably holding too closely to a fixed diurnal pattern.

Correlation related measures appear to be useful only as rough indicators of model performance. The use of the correlation coefficient needs to be discussed separately for daily maximum concentrations and for hourly concentrations. In analyzing daily maxima, the correlation should be used in conjunction with the slope and intercept of the best-fitting straight line, and should be accompanied by a scatterplot of C_o vs. C_p. The use of correlation alone would not reveal linear transformations or nonlinear relationships. Also, there are practical problems in the use of correlation and regression measures on a small set of extreme maxima. Data that might form the upper end of an acceptable straight line over a larger range of concentrations may appear to be an uncorrelated swarm of points when the range is severely restricted. Furthermore, on the small set of ozone maxima available for Denver, the standard errors on the regression coefficients were large indicating that the estimates were quite unreliable.

In analyzing hourly concentrations, the correlation is primarily a measure of the model's ability to replicate the average shape of the diurnal pattern of ozone concentrations. This is useful information, of course, for establishing confidence in the general performance of the model. But it does not measure the types of error that are most relevant to regulatory use of the model. An error which is of great concern for regulatory purposes, inaccurate prediction of the magnitude of the daily peak, may cause little or no reduction in the correlation. On the other hand, an error which is of less concern in regulation, missing the peak in time by an hour or two, will reduce the correlation. In comparison, hourly biases for each hour of the day provided more relevant information about the diurnal patterns for judging model performance and for diagnosing errors in the models. Time series plots of C_o and C_p provided more detailed diagnostic information for each day.

Spatial correlation was not computed in the Denver evaluation because five monitoring stations could not provide enough locational detail. To investigate spatial patterns, contour plots of the predicted concentrations were compared with observed concentrations at the five sites. This was useful in understanding spatial errors, which tended to be different for each day. As with the other performance measures, we suspect that the spatial correlation would only point out the days which have the largest spatial errors. Explanation of the errors would then be necessary and would require detailed investigation.

Comparison of the observed and predicted trends that result from changing emissions presents particular problems in a model evaluation. For the other performance measures, a reasonably accurate observed value was available for comparison with each predicted concentration. Unfortunately, an analogous observed trend caused solely by emissions change is not available for comparison with the predicted trend. It is necessary to use observed data that compounds the effects of emissions change and meteorological change. To separate the effects of the two changes requires either an accurate model of the meteorological effects or a model of the shape of the trend due to emissions change. In either of these cases, the errors in the resulting estimate of the "observed" trend may be rather large. Thus the predicted trend will be compared against a rather unreliable number, unless every effort is made to verify the modeling of the observed trend.

The linear trend approach used for the Denver evaluation was appropriate for Denver ozone concentrations in 1975-80. This was verified using vehicle travel counts from the Colorado Department of Highways and vehicle fleet emissions estimates from the federal mobile source emissions model. Such checking should be done before the same method is applied to similar model evaluations in other locations.

Evaluation of Graphical Displays

All of the graphs suggested earlier in this report were found to be useful in the example evaluation. In addition to the commonly used plots of observed and predicted values and residuals, graphs of hourly bias (with confidence intervals) and daily bias were helpful in interpreting the statistical measures.

Special graphical displays related to specific problems in air quality models were valuable for diagnosing causes of error. In the Denver evaluation, contour plots of the predicted concentrations were made for each hour of each day. These contour plots were important for understanding errors in the spatial location of the predictions. Contour plots of the emissions data used as input to the model were useful in finding locations where errors in emissions input could be causing errors in the predictions. In addition, wind trajectory plots were useful in tracking the development of the predicted ozone peak within the model.

This experience indicates that graphs should be an integral part of any performance evaluation. They go beyond the summary statistics in highlighting the type and location of errors in the predictions.

Evaluation of the Use of Subgroups of the Hourly Concentrations

Statistics which were averaged over the full hourly data set provided only very general information, merely an impression of high bias and low precision, with correlations that were high enough to indicate some success in capturing the diurnal cycle of ozone production. So many effects were averaged together that specific conclusions about the usefulness of the models were not possible.

Sorting the data by site provided the additional information that predictions at one site (CARIH) were substantially more biased than at the other sites, suggesting that special features of that site should be examined.

Sorting the data by day was important to determine whether any of the sample days presented particular difficulty to the models, in case some unusual atmospheric phenomena cannot be duplicated in the models. In the Denver data there was little difference between days in model performance when averaged over all sites, therefore no day required special analysis. But at specific sites, unusual model performance on particular days indicated by the daily bias and noise revealed specific site-related problems in the modeling or in the input to the models. Daily measures should not form the primary basis of an evaluation, however, because all of the information on the diurnal ozone pattern has been lost. Because of high autocorrelation over successive hours, confidence intervals on daily bias estimates will be extremely large.

Sorting the data by hour and by site was most useful for diagnosing errors in model performance. It revealed a systematic pattern of bias over the day at every site, with the models tending to overprediction in the early morning (7 - 9 a.m.) and to underpredict, with statistically significant bias, in the mid-day peak hours and the afternoon. Here, 95% confidence intervals on the hourly bias were especially useful, because they helped distinguish between bias which was a real, systematic feature of the model and bias which might be only the result of random fluctuations in the data. Sorting by hour and by site, both bias and noise measures averaged over all of the days helped to reveal patterns of error and led to diagnosis of a number of reasons for error in predictions at specific sites. Still smaller subsets, involving hourly averages for selected days at a particular site, were useful for diagnosing specific problems and for confirming hypotheses about the sources of particular errors.

Estimating the Bias in the Predicted Peak Concentration

The peak predictions are important in regulatory use of the model, therefore it is desirable to estimate the bias in those predictions. The bias estimate depends, however, on which predicted concentration is matched with the observed daily maximum. Theoretically it might be argued that the predicted daily maximum should be chosen only from a monitoring site location, because the chance of hitting the true area-wide peak would then be the same for both the observed and predicted maximum (as discussed in the section "Ways of Pairing Daily Maximum Concentrations"). That argument only holds if the days and the monitoring locations have been chosen randomly, however. In selecting days with high observed ozone concentrations and locating the monitoring stations in high-pollution areas, we have increased the chance of hitting the true area-wide peak in the observed concentrations. The chance of hitting the true peak in a prediction at a monitoring site has not been increased accordingly, however, because spatial errors in the predictions must be expected. The result is that, for a high ozone data set, this pairing should tend to produce some underprediction of the maximum, or positive bias.

The opposite tendency can be expected if the predicted maximum is chosen from the entire grid area covered by the model. In that case, the predicted maximum is chosen from a larger number of locations than the observed maximum. Thus this pairing should tend to produce an overprediction of the maximum, or negative bias.

We conclude that a meaningful statement of the bias in the predicted maximum should fall somewhere between the bias estimates produced under the two pairings just described. Where it falls in that range would depend on the population of days used in the evaluation, the locations of the monitoring sites, and the kinds of errors in the model. Errors in the spatial extent of the peak-ozone cloud, errors which systematically affect the prediction at a particular monitoring site, or errors which increase the likelihood of the predicted peak missing a monitoring site will each affect the bias in a different way. Furthermore, some real peaks may totally miss the monitoring sites and not be observed at all. To obtain the most accurate bias estimate, it may be necessary to match observed and predicted peaks by hand, using contour plots of the predicted concentrations. Thus establishing the bias in the peak predictions is not completely straightforward.

In the example evaluation, we conclude that the bias in EPA2's peak ozone predictions should be estimated between 10% and 31%, based on biases shown in Table 20. The best bias estimate within that range, and the most appropriate prediction of the daily maximum, will depend upon the types of spatial errors that are affecting the predictions. Spatial errors could not be adequately investigated in the Denver data set, thus further investigation is needed.

Problems in Comparing Models on Hourly Data

At first glance it is somewhat surprising that the EPA2 model looks distinctly better than the others in comparisons of the daily maxima, but not in comparisons of the complete set of hourly data. Two factors in the time-paired hourly data tend to mask the superior performance of EPA2 in predicting the peaks: missing the peak in time, and errors in the off-peak hours.

When the models miss the peak in time, predicting a maximum several hours before or after the observed maximum, the effect on the performance measures in a paired comparison can be quite misleading. An example of this is shown in Figure 38(a), the observed and predicted concentrations for a day at CARIH. All three models predict the peak 2 hours too late. The maximum predicted by the DOT model is much too low, but that predicted by EPA2 is acceptably close to the observed maximum. We would definitely select the EPA2 predictions as the superior ones in this example.

The performance measures would have suggested a different conclusion, however. For this day at CARIH they are

	DOT	EPA1	EPA2
Bias	29.3	21.9	21.3
Noise	31.4	34.9	40.9
Absolute deviation	31.0	30.6	34.1
r (C_o vs. C_p)	.68	.60	.47

Except for the bias in the DOT model, both DOT and EPA1 appear to be superior to EPA2. This judgment is based on their lower noise and absolute deviation, their higher correlation, and a comparable bias in EPA1. The apparent inferiority of EPA2 in this case results entirely from its superior estimation of the magnitude of the daily maximum. The three models perform similarly at other hours.

This is not an isolated case. The site maximum (i.e., the local peak at a given site for a given day) was predicted at the wrong hour by all three models more than 60% of the time, and by individual models even more frequently. This can be expected to have a substantial impact on statistical measures of model fit under hourly pairing, as shown above. If the goal is to find the model which best predicts the magnitude of the daily maximum, hourly pairing can lead to the wrong choice. Instead, in that case, the choice should be made by comparing model performances on the set of daily maxima, not paired by hour.

Another example in which the statistical measures are misleading is shown in Figure 38(b). This day at CAMP illustrates a problem in averaging hourly differences over all hours of a single day. Inspection of the observed and predicted concentrations again leads us to prefer EPA2 because it most closely approximates observed O_3 levels over several peak hours. The performance measures indicate otherwise, however, as shown below.

	DOT	EPA1	EPA2
Bias	1.3	-3.8	-8.1
Noise	17.2	14.5	14.7
Absolute deviation	12.1	11.6	13.8
r (C_o vs. C_p)	.85	.89	.88

EPA1 looks better than EPA2 on every measure, partly because the peak hours are shifted and partly because of large errors in EPA2 in the morning. By averaging over the day, a variety of model errors are merged in each measure, obscuring the most important distinctions between models. When statistics are averaged over several full days, there is even more tendency for error effects to balance out. It is no wonder, then, that our summary statistics on the full set of paired hourly data (Table 2) showed no clear distinctions between models and gave little information on the sources of errors in the models.

Effects of Non-normality on Bias Comparisons

The Kolmogorov-Smirnov test of normality was applied to all of the sets of daily- and site-maxima (observations, predictions, and residuals). In addition, it was applied to the separate hourly sets of EPA2 residuals, both for each site separately ($n = 11$) and for all sites together ($n = 55$). In no case could the hypothesis of normality be rejected, even with the rather liberal significance level of $\alpha = 20$. Such a result is conventionally taken to indicate that use of t- and F-tests is appropriate.

A test of normality on a sample containing only 11 points cannot be very precise, however. Therefore bias comparisons were done using both the t-test and the Wilcoxon Paired Rank Test to determine whether the Wilcoxon test could provide additional information.

On the hourly EPA2 residuals, the results of the two tests were virtually the same. When all sites were analyzed together, testing whether the bias was significantly different from zero at each hour, the two tests agreed on each of the 12 hourly data sets. Separating the data by site produced 60 sets of residuals, each containing approximately 11 points. Hypothesis-testing results on these 60 biases using the Wilcoxon and t-tests disagreed only twice, and the significance levels were close to the .05 borderline.

Most of the sets of data involving peak concentrations showed very large biases, thus it is not surprising to find that the Wilcoxon and t-tests agreed that these biases were significantly different from zero. The two tests agreed also on the one case in which bias was not significantly different from zero.

The two tests frequently did not agree, however, when residuals from two models were compared on peak concentrations. In these comparisons, the differences in bias were relatively small, hence the sensitivity of the test could be critical. In each case, the null hypothesis to be tested was that no difference exists between the bias in two models on a given set of peak concentrations. Such tests were performed on the two sets of site maxima ($n = 55$) and on the three differently matched sets of daily maxima (each having $n = 11$). In each data set, the Wilcoxon test found significant differences in bias which were not detected by the t-test. In no case did the opposite occur; that is, on the rare occasions when the null hypothesis was rejected by the t-test, the Wilcoxon test agreed.

We conclude that the nonparametric Wilcoxon test, with fewer restrictive assumptions then the t-test, can be more powerful than the t-test when the normality of the data is in doubt. This was shown to be true in spite of the fact that the data "passed" a Kolmogorov-Smirnov normality test.

In this comparison, Students' t was found to be overly conservative, less able than it should be to reject the null hypothesis at the specific confidence level. Nevertheless, it may still be useful to compute confidence intervals using Students' t. If this is done, it should be remembered that these intervals do not accurately represent the specified level of confidence. In this study it appears that the confidence intervals are unnecessarily large. In general, though, the type of error will depend upon the underlying distribution.

Uncertainty about the correctness of confidence intervals on the bias casts some doubt on the appropriateness of using such intervals to establish performance standards. At the very least, distributions of residuals should be examined. For example, skewness in these distributions could require asymmetrical confidence intervals to give fair treatment to positive and negative biases.

B. RECOMMENDATIONS

Recommended Performance Measures

The choice of performance measures should be based on the model attributes which are to be evaluated. Thus no single list of measures will be appropriate for every evaluation. Recommendations based on the Denver example evaluation should be relevant, however, for other time-dependent airshed models. The list of performance measures, graphic displays, and data combinations which appear to be most useful in the evaluation of a time-dependent urban airshed model are shown in Tables 25 and 26. In the Denver study it was found that the long list of performance measures recommended for evaluation of air quality models by the AMS workshop could be reduced considerably. Some of the recommended measures were simply redundant, as is the case with mean square error and absolute deviation. Others would be inappropriate, given the intended application of these models to a small number of selected high ozone days, or even a single worst-case day, in the evaluation of state air pollution control strategies. In particular, if the models are to focus on only a few days, comparison of frequency distributions of concentrations over long periods are not appropriate. Furthermore, both observations and predictions will be closely tied to the meteorological characteristics of the few chosen days, therefore pairing by day must be maintained in the comparisons.

The use of graphs to display results is to be encouraged at every stage of a model evaluation. Scatter diagrams, time series plots, and contour plots are essential aids in intepreting the statistics, and are useful, as well, in uncovering sources of error in the models.

Two types of peak comparisons which we tried have not been included in Table 25: the predicted daily maximum at the site of the observed maximum, and the predicted site maximum at the hour of the observed site maximum. These were less important than the other peak comparisons in this study because the information they offer about missing the peak in time and space was obtained from the hourly data. They do help by substantiating those findings, however, and they definitely should be included in a study that involved only peak concentrations and not hourly comparisons.

Before using statistical tests or confidence intervals, two assumptions underlying their use must be confronted: normality and independence. The effect of non-normality on the t-test was checked in some detail, and found to cause some significant differences in bias to be overlooked in our data. As a result, use of the Wilcoxon test to compare sample biases is strongly recommended when normality is in doubt. The assumption that the data is normally distributed is also important in use of the F distribution to compare variances, but deviations from normality have been empirically shown to have only a minor impact on this test. Therefore only in cases of extreme deviation from normality would one be unable to apply the recommended F-tests.

Lack of independence, due to autocorrelation of data in a time series, is a more serious problem. When air quality data is collected over successive hours or days some mutual dependence of data points is almost certain, and this dependence can seriously affect the accuracy of statistical estimates. Therefore autocorrelation should be measured and corrresponding adjustments made to the confidence intervals and degrees of freedom used in statistical testing.

One essential aspect of model performance has not been covered in earlier evaluations of the Urban Airshed Model, and no associated measure is included in the Workshop list. As the model is used in practice, it was important to evaluate the model's response to changing emissions. A simple comparison of linear trends in observed and predicted concentrations over several years was chosen, under constraints of limited data and resources. Independent information about Denver's vehicle travel and vehicle fleet emissions indicated that a linear trend was a reasonable assumption. In future research it would be worthwhile to try other approaches which are capable of better estimating the "observed" trend in ozone concentrations due to emissions changes over the years, controlling for changes in meteorology.

The Use of Statistical Measures as Performance Standards

Procedures recommended by Draper and Smith (1966) for evaluating regression models are widely used to determine whether a statistical model should be accepted or rejected. These methods primarily involve a careful study of histograms and scattergrams of the residuals from the fitted model.

Although some statistics are available for use in judging the acceptability of a model, in practice they are far less informative than the corresponding residual plots. For this reason, Draper and Smith do not recommend their use. More commonly used are statistical measures, such as R^2 and the F-test, for choosing the "best" of several models. Such measures formed the basis of the list of performance measures recommended by the AMS workshop. Even for choosing between models, however, Draper and Smith caution against the automatic use of statistics, saying, "sensible judgment is still required in the initial selection of variables and in the critical examination of the model through examination of residuals."

Performance of the three models examined here is not impressive by the usual standards applied to regression models, because these models have not been "fitted" to observed data. In addition to high bias and noise, a variety of systematic errors could be found in the residuals of each model. Yet these are "state of the art" models, the best available tool for scientific and regulatory analysis of the urban airshed. While we are forced to acknowledge that they are imperfect, it is still quite possible that they may be adequate for the purposes required of them.

Despite many attempts within the field of statistics to establish formal criteria of acceptability, statisticians emphasize the need for professional judgment based on the intended use of a model. We expect that it will be equally difficult (impossible?) to establish absolute criteria for deterministic models.

Comparisons and statistics like those listed in Tables 25 and 26 can provide, as suggested at the AMS/EPA workshop, a "rational framework for quantitatively evaluating the nature of differences between observations and prediction by models." However, this framework is likely to be skeletal. The performance measures provide "vital statistics" but not understanding. There are likely to be multiple causes of error in a model, some of which are serious for a given application and others of which are not. Diagnosis of the causes of error is necessary to determine their effects on regulatory applications. Then, judgment is needed to determine the seriousness of the errors and whether adjustments can be made.

Evaluating the Usefulness of a Model

Physical scientists and statistical modelers have, historically, approached modeling from fundamentally different points of view. The physical scientist tries to base a model as much as possible on underlying scientific truths which have been physically demonstrated in controlled experiments. Statistical models, on the other hand, are likely to be derived from observed data rather than from physical principles and can usually be validated only against limited samples of empirical data collected under relatively uncontrolled conditions. Thus statisticians are better able to accept the prospect of an imperfect model. The viewpoint of the statistician is well summarized by Phadke, Box, and Tiao.

> "On this view of modeling all models are wrong, but some models are useful. Thus while it is useless to seek a true one we can iterate towards successively more useful ones till we obtain one which is adequate for our purposes."

The user of air quality models looks for a "true" model because it is important to base decisions on the most scientifically correct available understanding of physical processes. But because these models are imperfect we, too, must iterate toward successively more useful ones until we obtain one which is adequate. From this perspective, the evaluation of a model is intimately connected to the objectives of its application.

The regulatory purpose requires accurate prediction of peak ozone concentrations on high ozone days, under changing emissions conditions. The change in emissions takes place gradually over a period of years, therefore a test of the models on data within a one or two year period is probably not sufficient. In this study, then, two separate tests were needed to match the regulatory purpose: 1) an analysis of daily maximum predictions under a variety of meteorological conditions represented by the 11 sample days in 1979-80, and 2) an analysis of predicted change in the daily maxima when emissions' input to the model was changed from 1979 to

1976 levels. In both of these analyses, EPA2 performed better than the other two models.

Although we have concluded that EPA2 is the best of the three models in the daily maximum predictions required for regulation, judgment will be required to determine whether that model performs well enough to satisfy the purposes of its users. Absolute performance standards can lead to the wrong decision, as shown in examples above. Even in comparing the performances of the models, knowledge of the nature of the differences in performance was necessary to choose between them. Statistical measures provide helpful initial comparisons and valuable clues for decision making, but they are not a substitute for detailed analysis upon which the decision making must depend.

REFERENCES

Barrett, J.P. and L. Goldsmith (1976), "When is N Sufficiently Large?", American Statistician, 30, pp. 67-70.

Brier, G.W. (1975), Statistical Questions Relating to the Validation of Air Quality Simulation Models, EPA-650/4-75-010, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 313 p.

Cole, H.S., C.F. Newberry, W. Cox, G.K. Moss, and D. Layland (1982a) "Application of the Airshed Model for Ozone Control in St. Louis," 82-20.1, 75th Annual Meeting of the Air Pollution Control Association, New Orleans, Louisiana, June 20-25, 1982.

Cole, H.W., W.M. Cox, D.E. Layland, G.K. Moss, C.F. Newberry (1982b) "The St. Louis Ozone Modeling Project," draft report, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 392 p.

Delaney, A. (1981) "The CHON Photochemistry of the Troposphere," Notes of the 1980 Summer Colloquium, Advanced Study Program and Atmospheric Chemistry and Aeronomy Division, National Center for Atmospheric Research, Boulder, Colorado, 172 p.

Demerjian, K.L., K.L. Schere and J.T. Peterson (1980) "Theoretical Estimates of Actinic (Spherically Integrated) Flux and Photolytic Rate Constants of Atmospheric Species in the Lower Troposphere," In Advances in Environmental Science and Technology, Volume 10, J. Pitts and R. Metcalf, eds., John Wiley & Sons, New York, New York, pp. 369-459.

Draper, N.R. and H. Smith (1966), Applied Regression Analysis, Wiley and Sons, New York.

Fox, Douglas G. (1981), "Judging Air Quality Model Performance," Bulletin American Meteorological Society, V. 62, No. 5, May 1981, pp. 599-609.

Greenberg, J. and P. Zimmerman (1982) private communication--unpublished data from 1980 Summer Colloquium, National Center for Atmospheric Research, Boulder, Colorado.

Haney, J.L., T.W. Tesche, and J.P. Killus (1983) "Application of the Systems Applications Airshed Model to the Philadelphia Metropolitan Area: 19 July 1979 Ozone Episode," U.S. Environmental Protection Agency, Contract No. 68-02-3582, Systems Applications, Inc., San Rafael, California, 1983, 121 p.

Hayes, S.R. (1979), Performance Measures and Standards for Air Quality Simulation Models, EPA-450/4-79-032. U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 313 p.

Hirtzel, C.S. and J.E. Quon (1981), "Estimated Precision of Autocorrelated Air Quality Measurements," Summaries of Conference Presentations, Environmetrics 81, pp. 200-201.

Hollander, M. and R.A. Wolfe (1973) Nonparametric Statistical Methods, John Wiley & Sons, New York, New York.

Keil, R. (1983) "The Impact of Meteorological Inputs on the Performance of an Urban Airshed Model," Masters Thesis, Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania (in press).

Kleiner, B. and T.E. Graedel (1980), "Exploratory Data Analysis in the Geophysical Sciences," Reviews of Geophysics and Space Physics, V. 18, No. 3, pp. 699-717.

Larsen, R.I. (1971), *A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards*, EPA Office of Air Programs Publ. No. AP-89, Research Triangle Park, NC, 56 p.

Layland, D.E. (1980) "Guideline for Applying the Airshed Model to Urban Areas," EPA-450/4-80-020, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 169 p.

Layland, D.E., S.D. Reynolds, H. Hogo and W.R. Oliver (1983) "Demonstration of Photochemical Grid Model Usage for Ozone Control Assessment," 83-31.6, 76th Annual Meeting of the Air Pollution Control Association, Atlanta, Georgia, June 19-24, 1983.

McRae, G.J. (1981) *Mathematical Modeling of Photochemical Air Pollution*, Ph.D. Thesis, Environmental Quality Laboratory, Report No. 18, California Institute of Technology, Pasadena, California, 754 p.

Myers, Jerome L. (1979), *Fundamentals of Experimental Design*, Allyn and Bacon, Inc., Boston (pp. 67-68).

Panofsky, H. (1981) Private communication.

Pearson, E.S. and H.O. Hartley (1976), *Biometrika Tables for Statisticians*, Vol. II, Biometrika Trust, London.

Phadke, M.S., G.E.P. Box, and G.C. Tiao (1977), Empirical-Mechanistic Modeling of Air Pollution." Proceedings of the 4th Symposium on Statistics and Environment, ASA, Washington, D.C. (pp. 91-100).

Reynolds, S.D., H. Hogo, W.R. Oliver, and L.E. Reid (1982) "Application of the SAI Airshed Model to the Tulsa Metropolitan Area," U.S. Environmental Protection Agency, Contract No. 68-02-3370, Systems Applications, Inc., San Rafael, California, 392 p.

Schere, K.L. (1982) "An Evaluation of Several Numerical Advection Schemes," draft report, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 37 p.

Tennekes, H. (1973) A model for the dynamics of the inversion above a convective boundary layer, J. Atmos. Sci., 30, pp. 558-567.

Tesche, T.W., C. Seigneur, L.E. Reid, P.M. Roth, W.R. Oliver, J.C. Cassmassi (1981) "The Sensitivity of Complex Photochemical Model Estimates to Detail in Input Information," EPA-450/4-81-031a, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 181 p.

Whitten, G.Z., J.P. Killus, and H. Hogo (1980) "Modeling of Simulated Photochemical Smog with Kinetic Mechanisms--Volume 1. Final Report," EPA 600/3-80-028a, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 348 p.

Table 1

Existence of High Pressure Influencing Denver

high pressure: ▨

no high pressure: ☐

Day	500 mb High Pressure Ridge				Surface High Pressure				Maximum Ozone on Modeled Day (ppb)
	Two Days Before	One Day Before	Modeled Day	One Day After	Two Days Before	One Day Before	Modeled Day	One Day After	
79180	▨	▨	▨	▨	▨	☐	▨	☐	153
79193	▨	▨	▨	▨	▨	☐	☐	☐	146
79208	▨	▨	▨	▨	▨	▨	▨	☐	162
79218	☐	▨	▨	▨	☐	☐	☐	☐	166
79249	▨	▨	▨	▨	▨	▨	☐	☐	157
80170	☐	▨	▨	▨	☐	☐	▨	☐	117
80177	▨	▨	▨	▨	▨	▨	▨	☐	117
80191	▨	▨	▨	▨	▨	▨	▨	▨	100
80204	▨	▨	▨	▨	☐	☐	▨	▨	154
80207	▨	▨	▨	▨	▨	▨	▨	▨	121
80219	▨	▨	▨	▨	▨	☐	▨	☐	101

Time is 1200 GMT equal to 0500 MST

Table 2

Meteorological Conditions I on High Ozone Days

Day	Max Temp* >80°F	Max Temp* >90°F	Daytime Precip.*	Wind Speed 1200 GMT 500 mb** (kts)	Wind Speed 1200 GMT Surface** (kts)	0600-1700 MST Maximum Wind Speed at Monitoring Sites (kts)***
79180	x			20	5	13
79193	x	x	T	15	5	20
79208	x	x		30	10	12
79218	x	x		10	5	10
79249	x			20	5	12
80170	x	x		25	5	12
80177	x	x	T	25	5	12
80191	x	x		35	10	12
80204	x	x		10	5	8
80207	x		T	30	5	17
80219	x	x		15	5	12

* at Stapleton International Airport (NWS)

** from daily weather maps

*** from Colorado Dept. of Health data

Table 3

Meteorological Conditions II on High Ozone Days

MST	Sky Cover (tenths)													Time of Maximum Temperature (MST)	Maximum Ozone on Modeled Day (ppb)
Day	5	6	7	8	9	10	11	12	13	14	15	16	17		
79180	0	0	0	0	0	0	2	3	3	4	5	8	5	1500	153
79193	1	0	0	2	2	3	3	4	5	5	9	9	9	1400	146
79208	1	0	7	8	3	3	3	4	5	9	9	5	8	1400	162
79218	0	0	0	0	0	0	1	1	0	0	0	0	0	1500	166
79249	0	0	0	0	0	0	0	0	0	1	1	0	0	1500	157
80170	0	0	1	0	0	1	4	4	6	4	6	7	4	1400	117
80177	0	0	0	0	0	0	0	0	1	2	3	9	9	1400	117
80191	0	0	0	0	0	1	2	5	3	2	2	5	5	1300	100
80204	10	7	4	0	0	0	0	0	0	1	3	5	6	1500	154
80207	0	0	0	0	0	0	0	2	2	2	10	10	4	1300	121
80219	0	0	0	0	0	0	0	0	1	2	5	9	6	1300	101

Note:

Insolation	Sky Cover
Strong	0-4
Moderate	5-7
Slight	7-8
Neutral	9-10

Table 4

Meteorological Conditions III on High Ozone Days

Day	Existence of Upper-Level Inversion below 2100 Meters at 1200 GMT	at 0000 GMT	Maximum Mixing Depth* (m)
79180	no	no	1900
79193	no	no	1900
79208	no	no	1450
79218	no	no	2100
79249	yes	no	2000
80170	no	no	2000
80177	no	no	4000
80191	no	no	2250
80204	no	no	2700
80207	no	no	1750
80219	no	no	2600

* As calculated by Tennekes' model

Table 5

Central Denver 5-Station Average Wind Speed (m/s)

(Department of Health Monitors)

Day	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	12-Hour Average
79180	1.7	2.1	2.6	1.9	1.8	2.2	2.1	2.2	2.5	2.5	4.2	4.2	2.50
79193	2.1	1.8	2.2	1.4	1.1	1.4	1.7	1.9	2.1	2.9	5.2	5.2	2.42
79208	1.4	2.0	1.8	1.2	1.1	1.1	1.6	3.0	2.1	2.6	3.2	3.6	2.06
79218	2.2	2.0	1.8	0.9	0.9	1.0	1.4	1.5	1.8	2.5	3.0	3.6	1.88
79249	1.4	1.6	1.8	1.4	1.2	1.1	1.9	1.8	3.0	3.7	4.1	4.1	2.26
80170	1.7	1.0	1.2	1.0	1.4	1.1	1.0	1.0	1.3	3.1	3.4	4.2	1.78
80177	1.8	1.7	1.4	0.9	1.5	2.3	3.0	4.3	4.1	2.5	1.6	1.8	2.24
80191	1.7	1.9	1.4	1.7	1.1	1.7	1.7	2.0	2.1	2.0	2.6	2.6	1.88
80204	1.8	1.4	1.0	1.3	1.5	0.9	1.5	2.8	3.3	2.7	2.6	2.3	1.93
80207	1.1	0.8	1.1	1.9	1.4	1.2	1.1	1.7	1.9	3.9	4.5	3.5	2.01
80219	1.5	1.3	1.1	1.7	1.5	2.7	2.7	3.4	4.3	2.9	4.5	3.3	2.58
11-Day Ave.	1.67	1.60	1.58	1.39	1.32	1.52	1.79	2.33	2.59	2.85	3.54	3.49	2.14

Table 6

Types of Wind Trajectories Occurring in the Six Hours Prior to the Observed and/or Predicted Ozone Peaks

Straight Through		Zigzag	Curved	Reversal
79180†		79193	79208	79218†§
	79249†	80170	80204*	
		80177	80207	
		80191		
		80219		

*Day with an approximately 3-hour "dead spot" in wind field at location of the predicted peak

†Days with maximum at Highland

§Day with highest daily maximum ozone of the 1977 and 1980 summers

Table 7

Sample Autocorrelation Function
(averaged over 5 stations and 11 days)

Data set	Autocorrelation functions lags 1 to 6	Estimates of ϕ from lags 1 to 3	Estimated ϕ	Effective n in 12 successive observations
Observed O_3	.77, .50, .32, .24, .20, .23	.77, .71, .68	.72	
DOT predictions	.78, .55, .44, .38, .34, .27	.78, .74, .76		
EPA1 predictions	.72, .42, .28, .25, .22, .18	.72, .64, .65	.67	
EPA2 predictions	.66, .32, .20, .18, .19, .17	.66, .57, .59	.61	
DOT residuals	.73, .45, .27, .20, .15, .13	.73, .67, .65	.68	2.9
EPA1 residuals	.70, .38, .22, .18, .14, .13	.70, .62, .60	.64	3.2
EPA2 residuals	.69, .37, .21, .18, .16, .13	.69, .61, .59	.63	3.3

Table 8

Performance Evaluation Measures for all Hourly Ozone Concentrations, Paired by Hour and by Station

Measure	Arvada	CAMP	CARIH	Highland	Welby	All Sites
$\bar{C}_o$	59.1	44.1	63.6	61.8	57.2	57.1
S_o	36.8	30.7	35.6	34.1	31.4	34.4
n	131	132	127	126	130	646
DOT Model						
Bias $\bar{d}$	11.1	2.7	25.0	9.7	13.9	12.4
Noise S_d	25.6	18.4	22.5	26.2	18.7	23.6
$\overline{\lvert d \rvert}$	20.5	12.8	26.2	17.0	17.5	18.8
95% conf., $\bar{d}$*	(2.2, 20.0)	(-3.7, 9.1)	(17.2,32.8)	(0.6, 18.8)	(7.4, 20.4)	(8.7, 16.1)
S_p	19.8	23.7	23.1	18.2	19.3	21.3
$r(C_o,C_p)$	.75	.80	.79	.65	.83	.74
EPA1 Model						
Bias $\bar{d}$	10.0	0.9	23.3	8.6	13.1	11.1
Noise S_d	24.3	18.4	22.7	24.0	18.5	22.8
$\overline{\lvert d \rvert}$	19.6	13.3	25.7	15.6	17.1	18.2
95% conf., $\bar{d}$*	(1.5, 18.5)	(-5.5, 7.3)	(15.4,31.2)	(0.2, 17.0)	(6.7, 19.5)	(7.6, 14.6)
S_p	21.1	24.8	24.7	19.5	20.1	22.5
$r(C_o,C_p)$	.78	.80	.78	.73	.83	.75
EPA2 Model						
Bias $\bar{d}$	11.8	-0.1	24.6	8.4	14.4	11.8
Noise S_d	25.1	18.7	24.1	25.0	19.1	23.9
$\overline{\lvert d \rvert}$	20.7	13.3	27.6	16.6	18.0	19.2
95% conf., $\bar{d}$*	(3.1, 20.5)	(-6.6, 6.4)	(16.2,33.0)	(-0.3,17.1)	(7.8, 21.0)	(8.1, 15.5)
S_p	20.4	25.9	25.5	19.9	18.9	22.7
$r(C_o,C_p)$	.76	.79	.74	.69	.82	.72

*Computed from $\bar{d} \pm 2S_e$ where $S_e = S_d / n_e$ and $n_e = 33$ for separate sites, $n_e = 165$ for all sites together, to adjust for autocorrelation in the differences.

Table 9

Bias and Noise for Each Hour Averaged over Eleven Days and Five Sites

Hour	DOT Model Bias	DOT Model Noise	EPA1 Model Bias	EPA1 Model Noise	EPA2 Model Bias	EPA2 Model Noise
6	1.9	8.5	1.9	8.5	1.8	8.2
7	.2	10.6	.2	10.4	-.2	10.6
8	2.9	14.2	2.5	14.0	2.4	14.7
9	5.1	16.9	3.9	16.8	3.1	18.6
10	10.3	20.5	8.4	20.2	8.6	21.9
11	17.2	27.0	14.4	25.5	14.5	26.1
12	26.1	28.7	22.9	28.1	22.7	28.7
13	21.6	24.7	19.6	23.2	21.9	24.1
14	18.7	28.6	16.6	28.7	18.0	31.0
15	17.8	30.1	16.2	29.7	18.5	31.0
16	13.7	23.1	13.3	22.4	15.1	22.8
17	13.3	22.0	13.5	21.6	14.6	21.4
Overall	12.4	23.6	11.1	22.8	11.8	23.9

Table 10

EPA2 and Highland: Differences Between Observed and Predicted Concentrations That Show Effect of Missing the Peak in Space

Day	Maximum Observed at Highland (ppb)	Maximum Predicted at Highland (ppb)	Concentration of Peak when Nearest Highland (ppb)	Maximum Concentration of Peak Passing by Highland (ppb)
79180	153	103	107	123
79218	166	82	105	132
79249	157	73	111	116
		Avg.Diff.=73 ppb	Avg.Diff.=51 ppb	Avg.Diff.=35 ppb

Table 11

EPA2 at Highland: Bias and Noise for Each Hour for Full Eleven Days and Eight-Day Subset

Hour	11-Day Sample Bias	11-Day Sample Noise	Eight-Day Subset** Bias	Eight-Day Subset** Noise	$\overline{C}_o$
6	8.2*	6.8	6.7*	6.7	22.0
7	-9.4*	6.6	-11.1*	6.3	19.5
8	-7.2*	6.3	-9.1*	5.6	29.6
9	-3.6	9.2	-8.1*	4.5	37.7
10	.2	13.7	-5.7	10.9	46.7
11	6.2	17.5	-2.4	9.2	56.8
12	9.3	24.2	.1	19.1	70.4
13	20.5*	27.8	6.1	13.2	73.3
14	25.7	41.0	3.0	13.3	64.3
15	20.3	44.4	-3.6	17.3	59.8
16	14.4	22.2	3.1	7.3	57.6
17	11.5*	14.8	8.3	10.1	57.3
Overall	8.4	25.0	-0.8	12.4	

*Bias is significantly different from zero at the 95% confidence level.

**The eight days analyzed separately at Highland are 79193, 79208, 80170, 80177, 80191, 80204, 80207, 80219.

Table 12

EPA2 at Arvada: Bias and Noise for Each Hour for Full Eleven Days and Eight-Day Subset

Hour	11-Day Sample		Eight-Day Subset**	
	Bias	Noise	Bias	Noise
6	-2.8	9.2	-5.3	7.7
7	-4.8	10.1	-6.4	8.4
8	-4.5	15.6	-7.9	17.2
9	-1.6	19.0	-7.1	17.9
10	6.4	19.7	.4	16.5
11	14.0	18.8	10.1	20.6
12	25.4	26.8	18.0	24.8
13	26.0	30.6	13.5	25.3
14	27.0	30.4	16.8	26.4
15	24.9	28.8	20.0	30.9
16	17.3	24.0	17.9	26.0
17	13.3	28.0	13.0	26.9
Overall	11.8	25.1	7.1	23.3

**The eight days analyzed separately at Arvada are 79208, 79218, 79249, 80170, 80177, 80191, 80207, 80219.

Table 13

Observed and Predicted Site Maxima at CARIH Monitoring Station

	CARIH Site Maximum* (ppb)			
Date	Observed	DOT	EPA1	EPA2
79-180	112	70	73	77
79-193	97	70	87	85
79-208	162	64	68	75
79-218	140	78	109	128
79-249	115	66	72	74
80-170	117	77	73	75
80-177	117	53	61	57
80-191	100	75	70	59
80-204	151	82	84	96
80-207	121	68	79	70
80-219	101	59	62	75

*Unpaired by hour

Table 14

Point Source Influence in EPA2:
Hourly Bias, With and Without Point Sources, at Each Site
Averaged Over Six Days*

Hour	Arvada		CAMP		CARIH		Highland		Welby	
	With	Without	With	Without	With	Without	With	Without	With	Without
6	-5.0	-5.2	2.3	2.3	8.3	8.3	8.4	8.4	-1.0	-1.0
7	-5.8	-5.8	.7	.3	14.8	14.7	-10.4	-10.6	.5	2.5
8	-3.0	-3.3	-2.3	-3.5	27.2	25.8	-8.8	-10.2	5.5	5.3
9	1.8	.2	-4.5	-6.0	33.3	26.8	-7.8	-11.6	7.5	4.5
10	9.2	6.5	4.0	-0.5	32.8	25.8	-7.0	-12.0	9.0	7.8
11	14.0	9.7	13.5	8.7	32.3	30.3	-1.0	-6.6	13.3	13.5
12	23.7	21.2	20.5	19.0	51.8	54.8	5.0	3.5	34.7	34.7
13	25.5	24.3	6.5	4.0	36.4	40.2	6.0	3.2	22.2	23.2
14	25.0	24.8	-6.3	-9.0	15.6	16.2	1.5	1.5	18.2	17.8
15	16.2	16.2	-10.3	-12.7	26.2	20.8	.2	-1.3	15.0	14.3
16	3.5	2.8	-11.8	-13.2	19.5	14.8	2.2	2.3	13.7	13.2
17	-1.0	-1.7	-6.8	-7.8	22.3	17.7	5.2	4.3	15.8	15.2
Overall	8.9	7.7	0.4	-1.5	26.2	24.0	-0.2	-2.0	12.7	12.4

*The six days analyzed for point source influence are 79193, 79208, 80170, 80191, 80204, and 80219.

Table 15

Point Source Influence in EPA2: Site Maximum Ozone Concentration, With and Without Point Sources.

Day	Arvada			CAMP			CARIH		
	C_o^{max} (ppb)	With C_p^{max} (ppb)	Without C_p^{max} (ppb)	C_o^{max} (ppb)	With C_p^{max} (ppb)	Without C_p^{max} (ppb)	C_o^{max} (ppb)	With C_p^{max} (ppb)	Without C_p^{max} (ppb)
79193	146	105	108	98	88	100	*	85	106
79208	120	72	78	121	73	73	162	75	72
80170	62	72	74	98	91	85	117	75	67
80191	54	58	59	76	68	70	100	59	63
80204	154	80	85	117	115	112	151	96	88
80219	90	73	76	81	71	68	101	75	75

*Missing Data during peak

Table 16

Background Sensitivity: Results II
Using EPA1

Daily Maximum Predicted Ozone

Day*	"Normal" Background			Low Background [20 ppb]			High Background [90 ppb]		
	Background Input (ppb)	Time of Max	Value (ppb)	Time of Max	Value (ppb)	% Change	Time of Max	Value (ppb)	% Change
79180	50	14	119	13	80	(-33%)	13	127	(+ 7%)
79193	55	12	113	12	100	(-12%)	12	131	(+16%)
79218	50	14	129	14	115	(-11%)	14	153	(+19%)
79249	45	14	109	14	97	(-11%)	14	138	(+27%)
80170	55	15	93	13	70	(-25%)	17	138	(+48%)
80191	50	13	85	13	65	(-24%)	11	138	(+68%)
80219	50	13	82	11	66	(-20%)	15	120	(+46%)
			$\bar{x}$ = 104.3		$\bar{x}$ = 84.7	(-19%)		$\bar{x}$ = 135.0	(+29%)
			s = 17.9		s = 19.5			s = 10.5	

*7 days picked at random from the full set of 11.

Table 17

Influence of Vertical Mixing in EPA1 for Days 79180 and 80204

	Base Run	Decreased Vertical Mixing Run	Percent Change
Daily Bias			
79180	16.9	17.5	3.6%
80204	14.3	13.9	-2.8
$C_o^{max} - C_p^{max}$, unpaired by site			
79180	60	41	-31.7
80204	70	38	-45.7
C_p^{max} over monitoring sites			
70180	93	112	20.4
80204	84	116	38.1
C_p^{max}			
79180	119	148	24.4
80204	105	141	34.3

Table 18

Neutral Day Sensitivity
Hourly O_3 Bias

ALL SITES

EPA1

HOUR	79180						80204					
	Base Run Simulation			Neutral Day Simulation			Base Run Simulation			Neutral Day Simulation		
	C_o (ppb)	C_p (ppb)	$\bar{d}$ $(C_o - C_p)$	C_o (ppb)	C_p (ppb)	$\bar{d}$ $(C_o - C_p)$	C_o (ppb)	C_p (ppb)	$\bar{d}$ $(C_o - C_p)$	C_o (ppb)	C_p (ppb)	$\bar{d}$ $(C_o - C_p)$
6			5.8			5.8			16.4			16.4
7			9.0			9.0			10.8			10.8
8			7.8			7.8			7.0			11.0
9			1.8			6.8			2.4			6.4
10			5.0			11.8			1.2			-0.8
11			9.8			10.4			18.6			7.8
12			14.8			11.2			44.6			28.0
13			25.2			22.0			34.6			28.8
14			30.6			28.8			14.6			19.6
15			39.2			38.0			6.6			10.8
16			20.8			22.6			6.2			14.2
17			32.6			35.2			9.0			14.0
Daily Bias			16.9			17.5			14.3			13.9
All-Station Maximum Difference Unpaired in Site			60			41			70			38

Table 19

Maxima for Each Day at Each Site (Site Maxima): Performance Measures with and without Time Pairing

Number of observed site maxima (5 sites, 11 days), n=55

Mean, $\overline{C^{max}_{o,s}} = 103.8$ Standard deviation, $S_o = 27.2$

Pairing and Model	Bias $\bar{d}$	95% conf. interval on bias estimate based on Student's t	Variability S_p	$F = \frac{S_o^2}{S_p^2}$	Noise S_d	Correlation r_{op}
Paired by hour: Predicted concentration at time of observed site maximum						
DOT	42.8	(35.0, 50.6)	11.1	6.00*	28.9	.055
EPA1	40.3	(33.0, 47.6)	11.6	5.50*	27.1	.225
EPA2	41.3	(33.8, 48.8)	14.9	3.33*	27.6	.248
Unpaired by hour: Predicted maximum over all hours, for given day and site						
DOT	34.2	(26.7, 41.7)	10.7	6.46*	27.6	.159
EPA1	29.3	(22.3, 36.4)	14.2	3.67*	26.0	.346**
EPA2	28.0	(20.8, 35.3)	18.0	2.28*	26.8	.354**

*All of the prediction variances are significantly smaller than the variance in the observed data. (Critical $F_{.05} = 1.60$ with d.f. = 54 for each data set.)

**Correlations between observed and predicted site maxima are significantly greater than zero. (Critical $r_{.05} = .273$ with d.f. = 54.)

Table 20

Daily Maximum Predictions:
Performance Measures for Different Pairing Methods

Number of observed daily maximum concentrations, n=11

Mean, $\overline{C_o^{max}} = 135.8$ Standard deviation, $S_o = 24.9$

Pairing and Model	Bias $\bar{d}$	95% conf. interval on bias estimate based on Student's t	Variability S_p	$F = \frac{S_o^2}{S_p^2}$	Noise S_d	Correlation r_{op}
(a) Paired by site: Predicted maximum at site of observed maximum						
DOT	66.2	(49.5, 82.8)	9.4	7.02*	24.8	.203
EPA1	59.2	(45.0, 73.4)	11.9	4.38*	21.1	.532
EPA2	58.2	(44.1, 72.3)	15.1	2.72	21.0	.539
(b) Unpaired by site: Predicted maximum over all monitoring sites						
DOT	57.6	(42.3, 72.9)	11.3	4.86*	22.8	.406
EPA1	50.6	(37.3, 64.0)	15.3	2.65	19.9	.602**
EPA2	42.7	(30.0, 55.4)	21.7	1.32	18.9	.678**
(c) Unconstrained in space: Predicted maximum over full modeled region						
DOT	40.7	(23.5, 57.9)	13.2	3.56*	25.6	.212
EPA1	30.3	(15.8, 44.8)	19.6	1.61	21.6	.550
EPA2	13.7	(-0.5, 27.9)	23.4	1.13	21.1	.620**

*Variance of predictions is significantly smaller than the variance in the observed data. (Critical $F_{.05}$ = 2.98 with d.f. = 10 for each data set.)

**Correlations between observed and predicted daily maxima are significantly greater than zero. (Critical $r_{.05}$ = .576 with d.f. = 10.

Table 21

Wilcoxon Paired Rank Tests
Comparing Daily Maximum Predictions by the Three Models

Comparison	# of non-zero differences	Rank sum T^-	Significance of T^-
(a) Paired by site: Predicted maximum at site of observed maximum			
Max. concentrations			
Observed vs. DOT	11	0	<.01
Observed vs. EPA1	11	0	<.01
Observed vs. EPA2	11	0	<.01
Residuals			
DOT vs. EPA1	11	8	<.05
DOT vs. EPA2	11	9	<.05
EPA1 vs. EPA2	10	23.5	--
(b) Unpaired by site: Predicted maximum over all monitoring sites			
Max. concentrations			
Observed vs. DOT	11	0	<.01
Observed vs. EPA1	11	0	<.01
Observed vs. EPA2	11	0	<.01
Residuals			
DOT vs. EPA1	11	12	--
DOT vs. EPA2	11	6	<.05
EPA1 vs. EPA2	11	7.5	<.05
(c) Unconstrained in space: Predicted maximum over full modeled region			
Max. concentrations			
Observed vs. DOT	11	0	<.01
Observed vs. EPA1	11	3	<.01
Observed vs. EPA2	11	12.5	--
Residuals			
DOT vs. EPA1	10	7	<.05
DOT vs. EPA2	11	1.5	<.01
EPA1 vs. EPA2	11	0	<.01

Table 22

Daily Maximum Predictions: Regression Against Observed Maxima

$$C_o = a + bC_p$$

Pairing and Model	Slope b	Intercept a	r^2
(a) Paired by Site			
DOT	.54	98.	.041
EPA1	1.11	51.	.283
EPA2	.89	67.*	.291
(b) Unpaired by Site			
DOT	.90	65.	.165
EPA1	.98*	53.	.362
EPA2	.78*	63.*	.460
(c) Unconstrained in Space			
DOT	.40	98.	.045
EPA1	.70	62.	.303
EPA2	.66*	55.	.384

*Significantly different from zero at the 95% confidence level. (Note that a "good" model should produce a slope which is significantly greater than zero and an intercept which is not significantly different from zero.)

Table 23

Daily Maximum Ozone Over the Full Grid Area Predicted Using 1979 and 1976 Emissions Inventories

Meteorology from Day	DOT Model			EPA1 Model			EPA2 Model		
	1979	1976	% Change	1979	1976	% Change	1979	1976	% Change
79180	106	115	8.5%	119	126	5.9%	123	131	6.5%
79193	87	91	4.6	113	120	6.2	123	130	5.7
79208	81	86	6.2	102	113	10.8	136	143	5.1
79218	107	114	6.5	129	136	5.4	133	150	12.8
80170	99	106	7.1	93	101	8.6	102	115	12.7
80191	94	104	10.6	85	96	12.9	91	105	15.4
80204	107	119	11.2	105	116	10.5	164	187	14.0
80219	82	85	3.7	82	84	2.4	95	111	16.8
Mean	95	102	7.4%	104	112	7.8%	121	134	10.7%

Table 24

Change in Ozone Resulting from Changes in Emissions: Estimated by Trend Analysis

	Estimated Avg. Annual Change (ppb)	Average daily max. O_3 on 8 Sample Days in 1979-80	Annual Changes as % of O_3 on the 8 Sample Days	Standard Error in Annual % Change	3-year Change as % of O_3 on the 8 Sample Days
Observed O_3, 1975-80	-6.20	137.	4.5%	1.0%	13.6%
EPA2 Prediction					
Maxima at sites	-3.46	98.*	3.5	3.4	10.6
Maxima over grid	-4.38	121.*	3.6	3.5	10.9
EPA1 Prediction					
Maxima at sites	-2.50	87.5*	2.9	2.8	8.6
Maxima over grid	-2.67	103.5*	2.6	2.7	7.7
DOT Prediction					
Maxima at sites	-1.75	82.*	2.1	2.1	6.4
Maxima over grid	-2.38	95.*	2.5	2.2	7.5

*Predicted O_3 based on 1979 emissions inventory.

Table 25

Recommended Combinations of C_o and C_p

For evaluating accuracy of the peak prediction:

1) C_o^{max} (s) with C_p^{max} (x) compares max. obs. and max. pred. for each day, where the predicted max. is constrained to be at the location of a monitoring site. (1 set of statistics)

2) C_o^{max} (s) with C_p^{max} (g) compares max. obs. and max. pred. for each day, where the predicted max. is selected from any grid point in the modeled region (1 set of statistics)

For diagnosis of site-specific daily maximum problems:

3) C_o^{max} (s,h) with C_p^{max} (s,x) compares max. obs. and max. pred. for each day at a given site, unpaired by hour. (A set of statistics for each site, plus another averaged over all sites.)

For diagnosis of sources of error:

C_o (s,t) with C_p (s,t) compares obs. and pred. concentrations matched by site and time, with the data sorted in the following ways:

1) By site: One set of statistics for each site, averaged over all hours and all days.

2) By day and site: One set of statistics for each day/site combination, plus one set for each day averaged over all sites.

3) By hour and site: One set of statistics for each hour/site combination, plus one set for each hour averaged over all sites.

Table 26

Recommended Statistical Estimators and Displays

Statistical Estimators:

Bias $\bar{d}$, with confidence interval based on a one-sample (paired) t with adjustments for autocorrelation. Bias comparisons based on the Wilcoxon paired rank test should also be done if the data sets are suspected not to be normally distributed.

Noise S_d; noise from two models can be compared using an F-test.

Absolute deviation $\overline{|d|}$, as measure of gross error.

Unpaired variability comparison $F = \frac{S^2_{C_o}}{S^2_{C_p}}$, with significance tested using the F-distribution.

Correlation, slope, and intercept from simple linear regression of C_p vs. C_o.

Graphical Displays of Data and Statistics

Plots of observed and predicted concentrations vs. time, and of observed and predicted hourly means vs. hour of the day.

Scatterplots of C_p vs. C_o, differences vs. C_o, and differences vs. hour of the day.

Plots of bias vs. hour and bias vs. day, with confidence intervals.

Contour plots of predicted concentrations and of relevant model inputs.

Any other plots and data combinations that might illuminate particular aspects of the model.

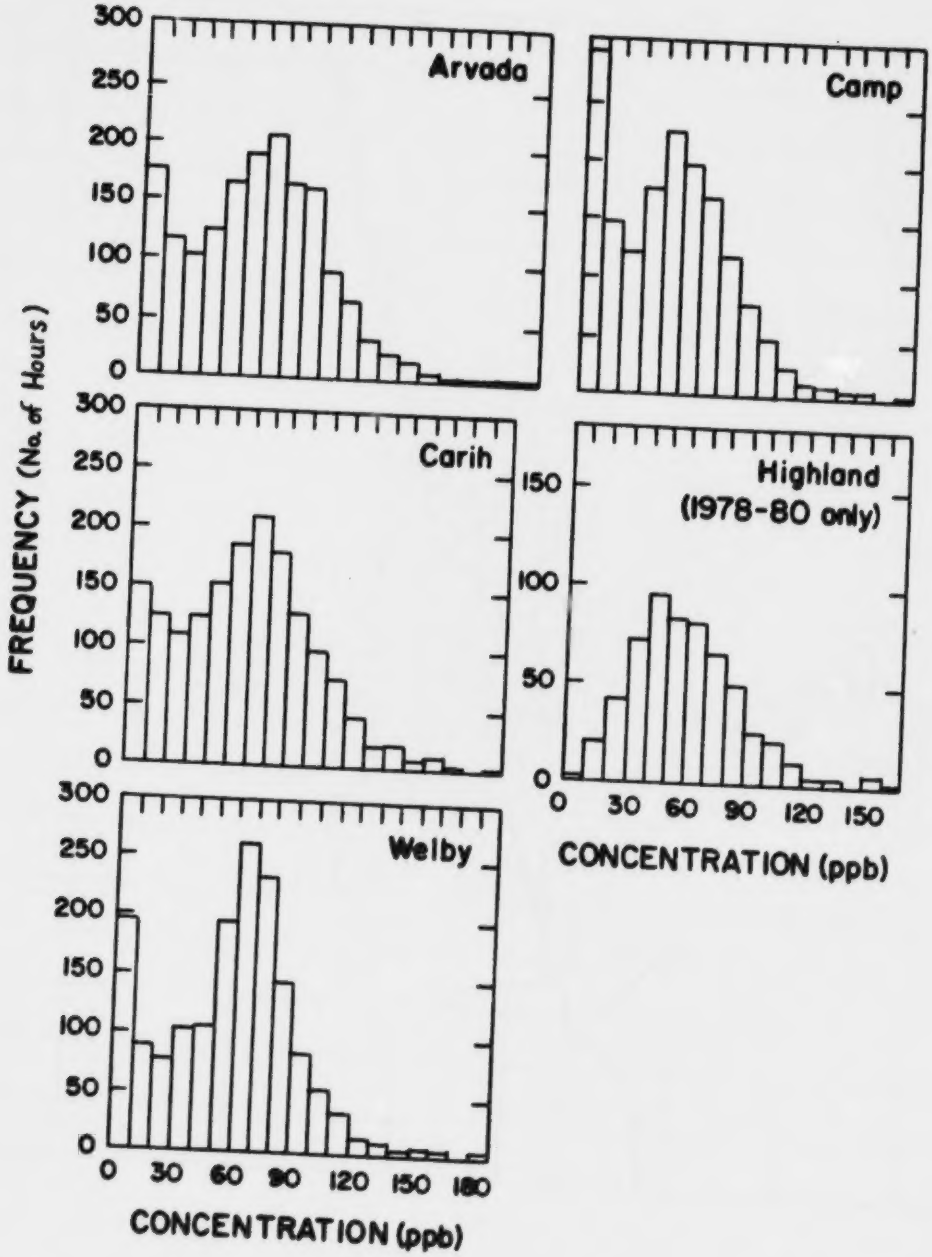


FIGURE 1: Distribution of Hourly Ozone Concentrations on High-Ozone Weekdays, May-September, 1975-1980 (≥.100 ppm)

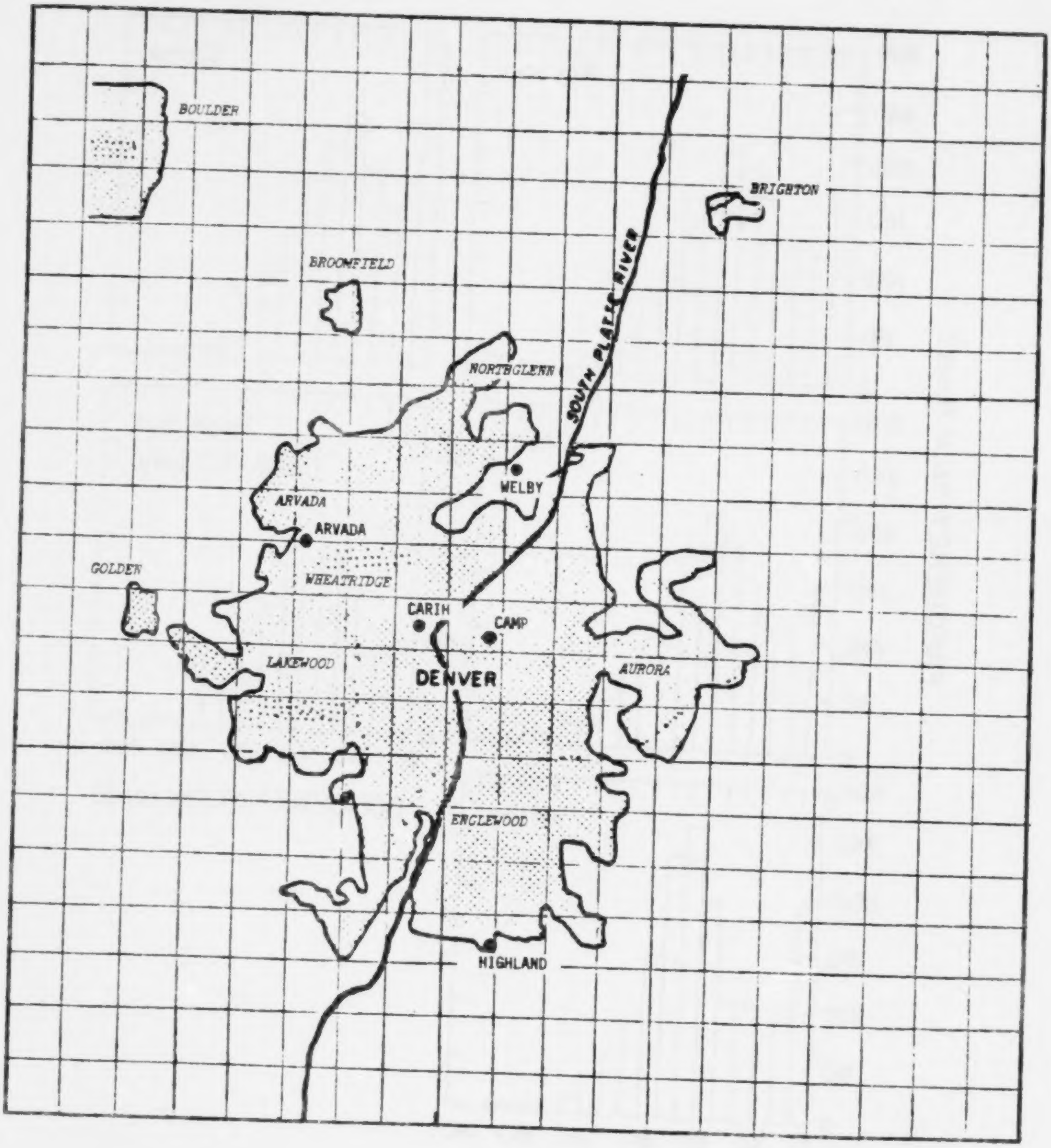

BOULDER
BRIGHTON
BROOMFIELD
SOUTH PLATTE RIVER
NORTHGLENN
WELBY
ARVADA
ARVADA
GOLDEN
WHEATRIDGE
CARIH
CAMP
LAKEWOOD
DENVER
AURORA
ENGLEWOOD
HIGHLAND

FIGURE 2: Denver Modeling Region Showing the 5 Monitoring Stations, Arvada, CAMP, CARIH, Highland and Welby, and the Model Grid.

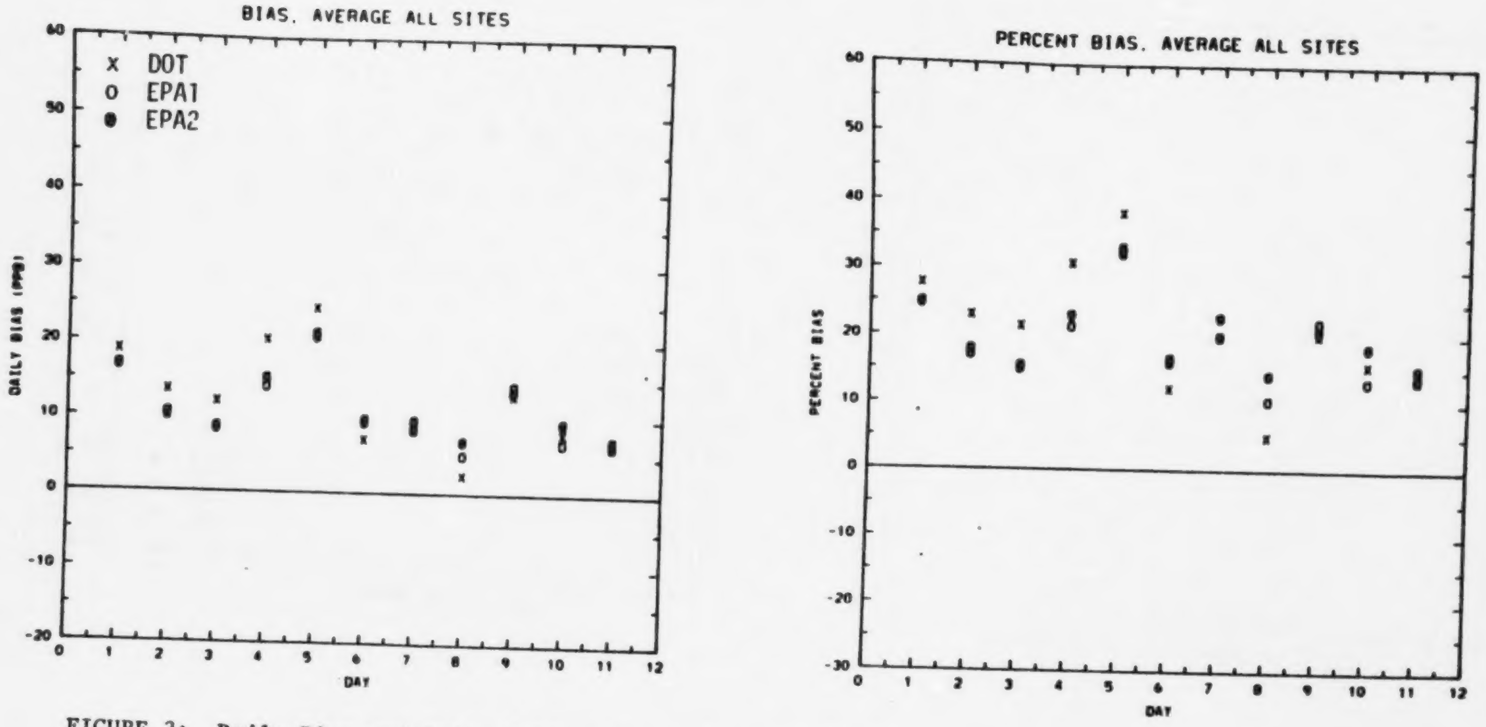


FIGURE 3: Daily Bias and Daily % Bias for Each of the 11 Days averaged over the 5 Monitoring Sites

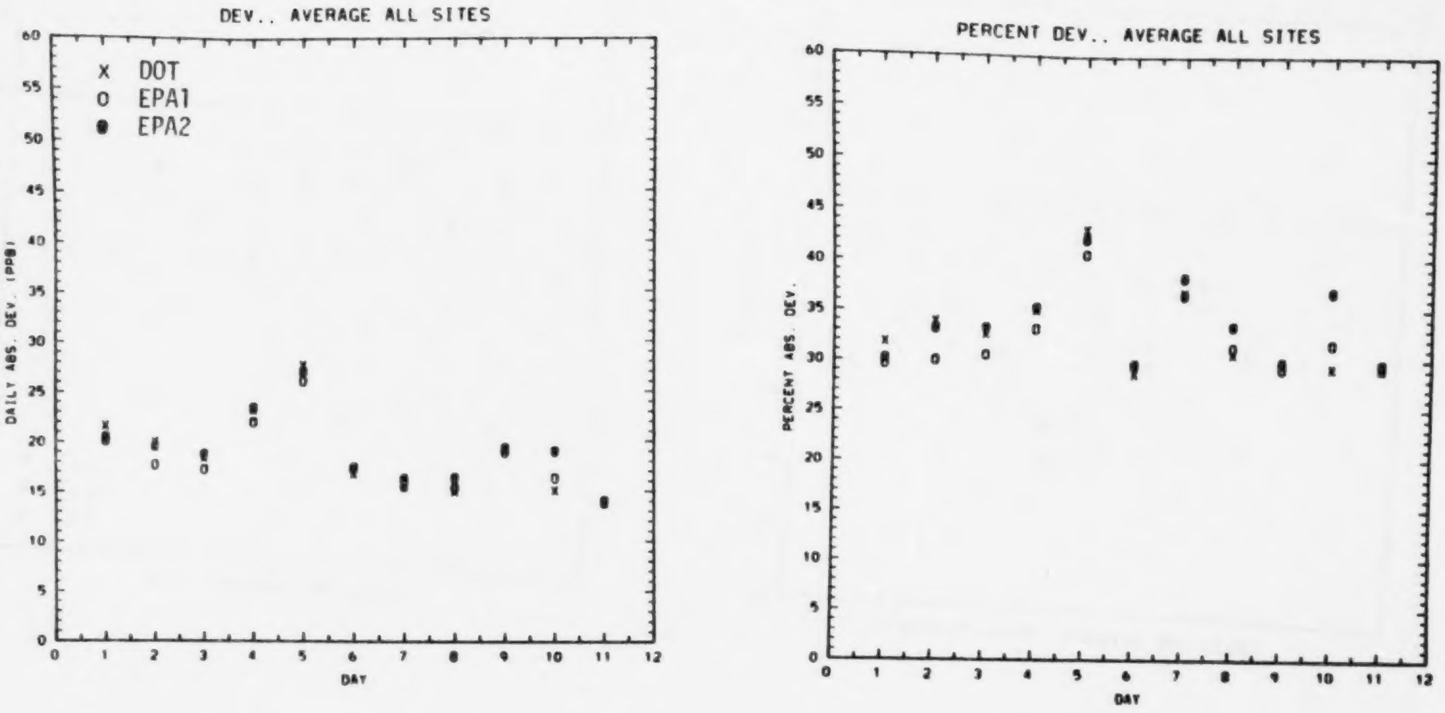


Figure 4: Daily Absolute Deviation and Daily % Absolute Deviation for Each of the 11 Days Averaged over the 5 Monitoring Sites.

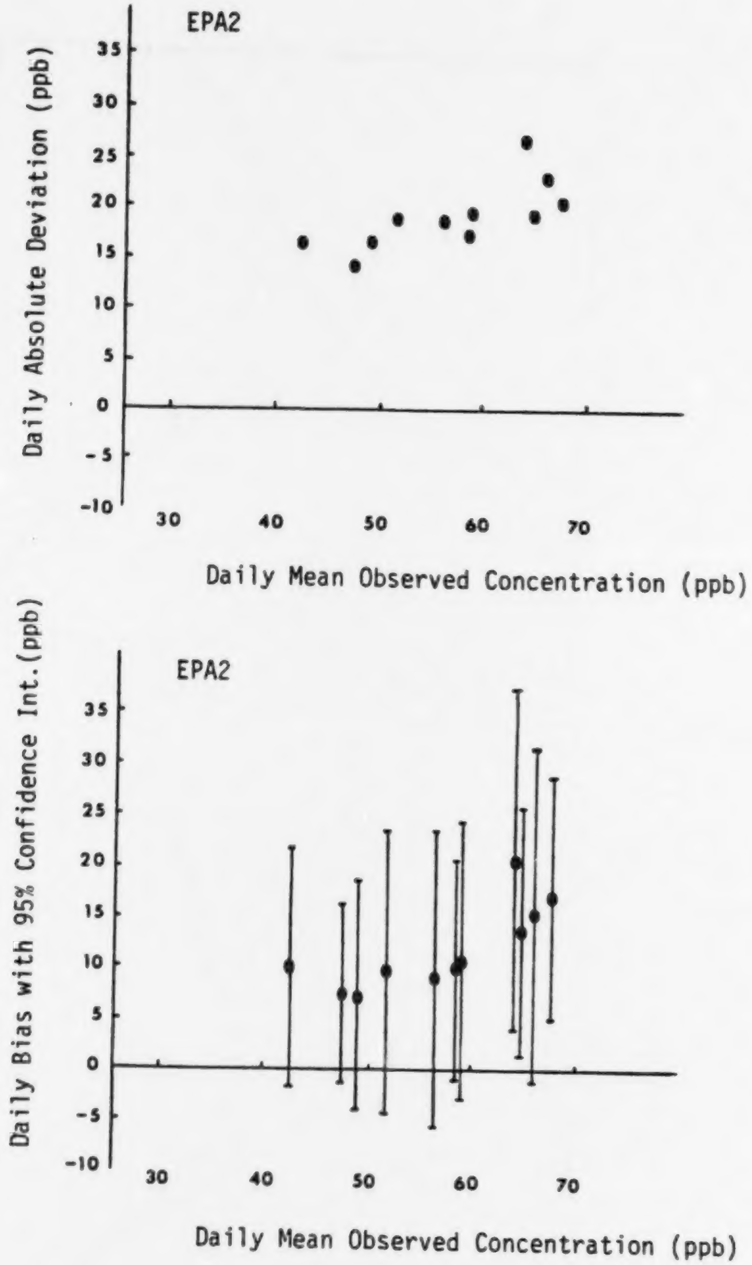


FIGURE 5: Daily Absolute Deviation and Bias in EPA2 versus Mean Observed Concentration, Averaged Over Five Sites.

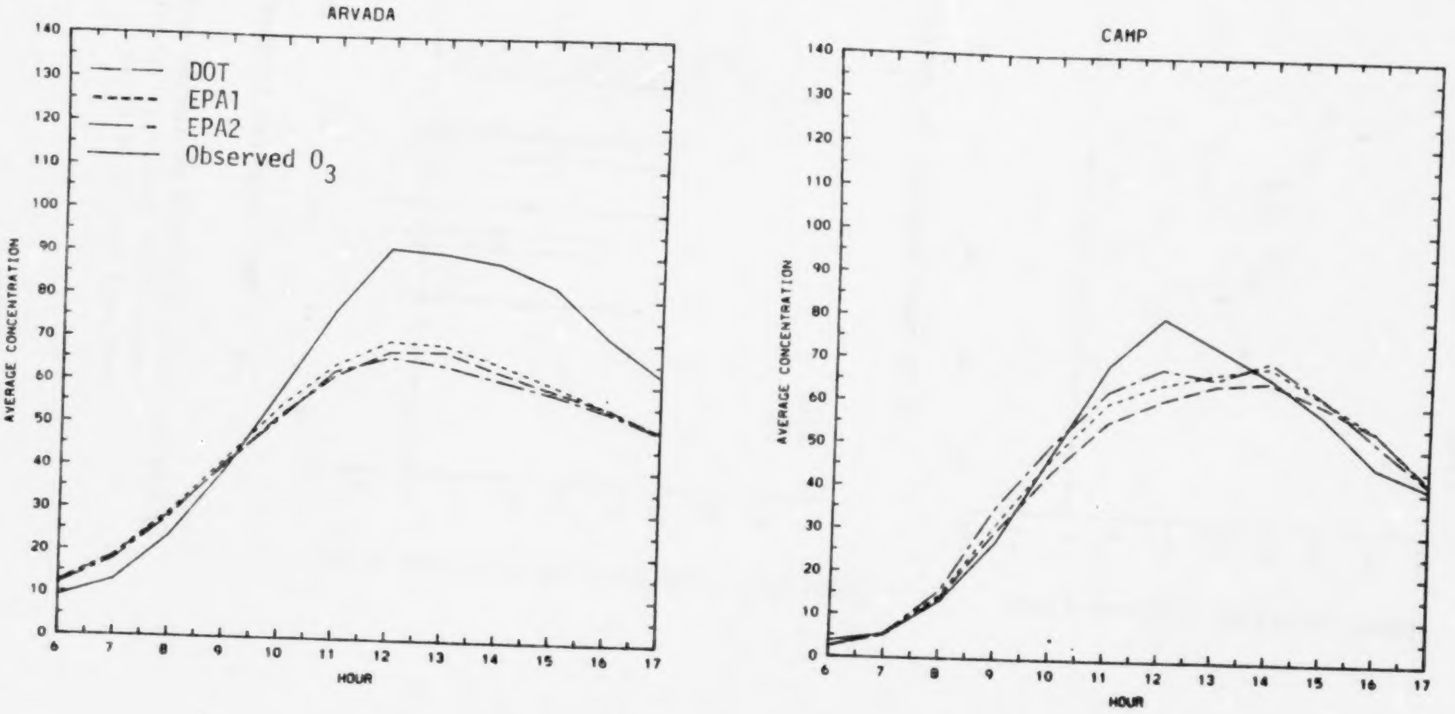


FIGURE 6: Observed and Predicted Diurnal Patterns at the 5 Monitoring Sites Averaged Over the 11 Days.

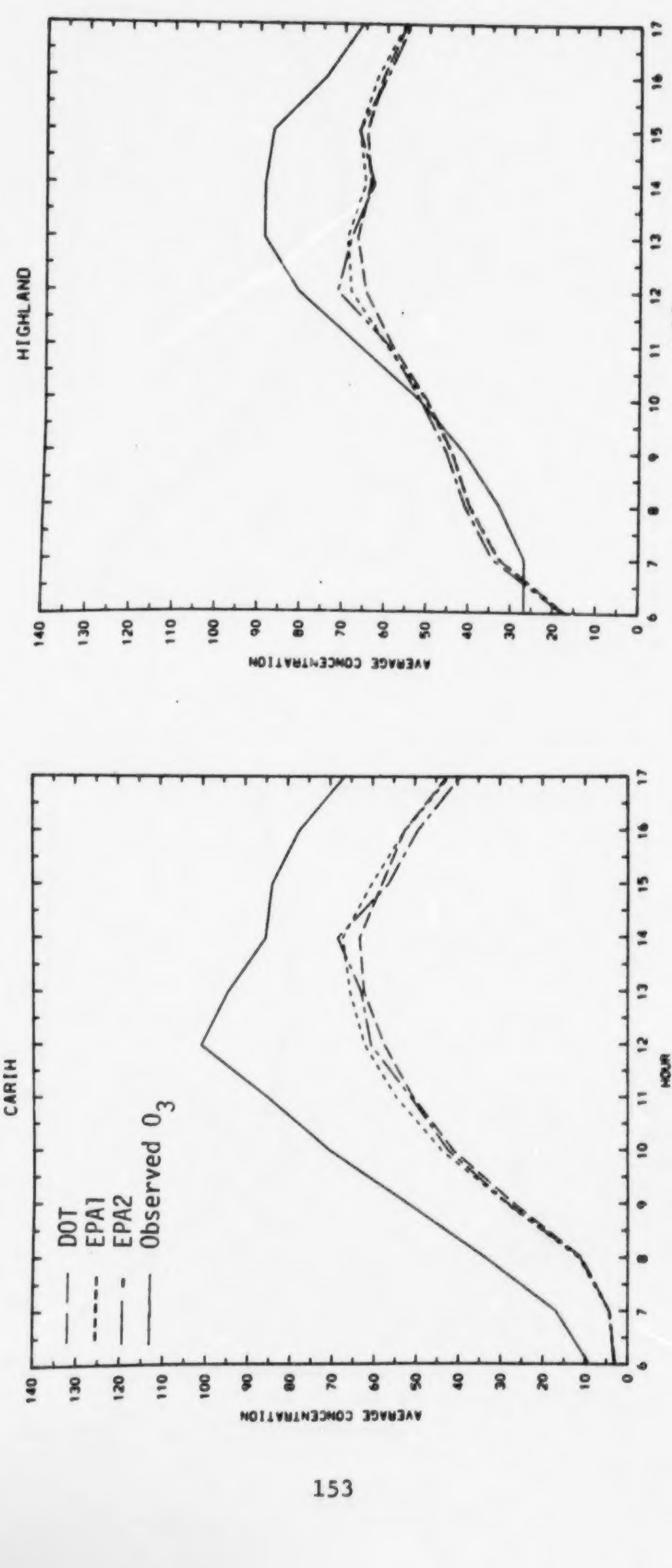


FIGURE 6, (Continued)

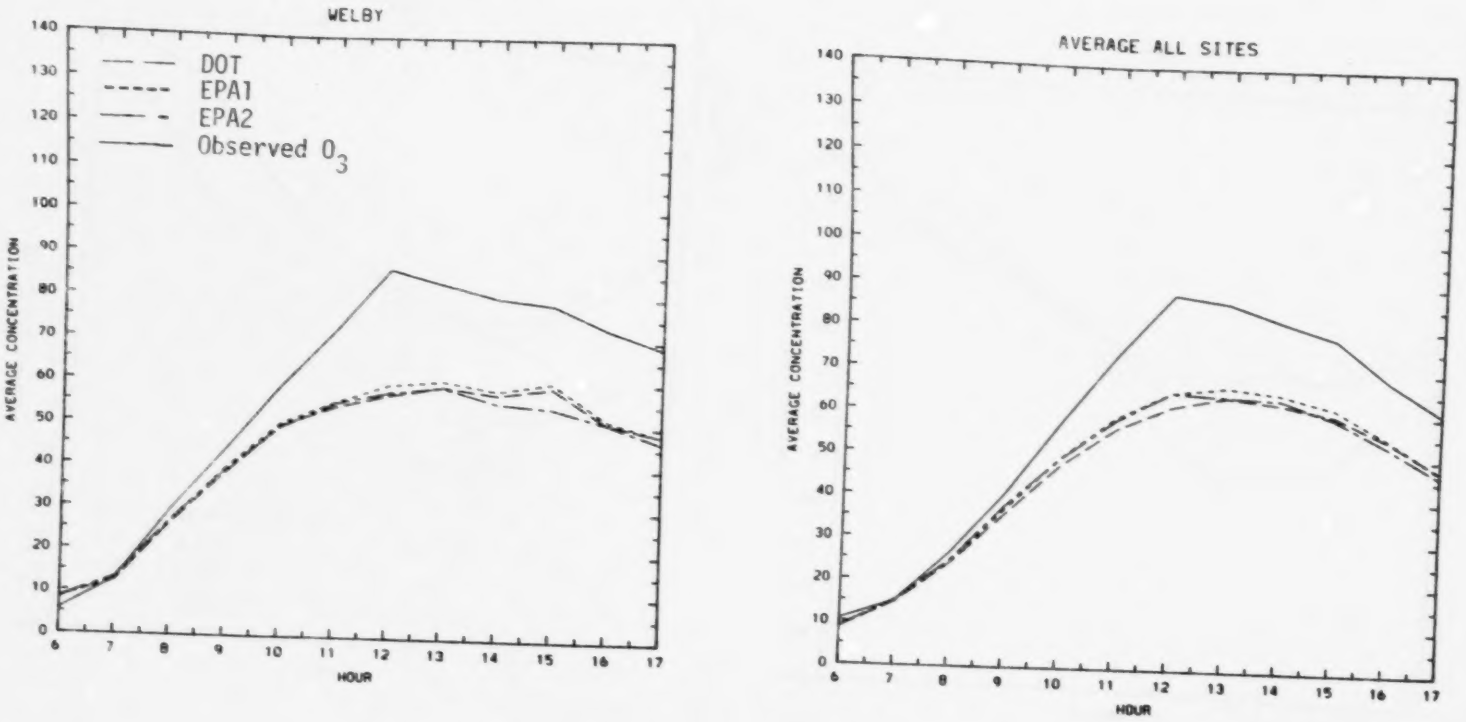


FIGURE 6, (Continued)

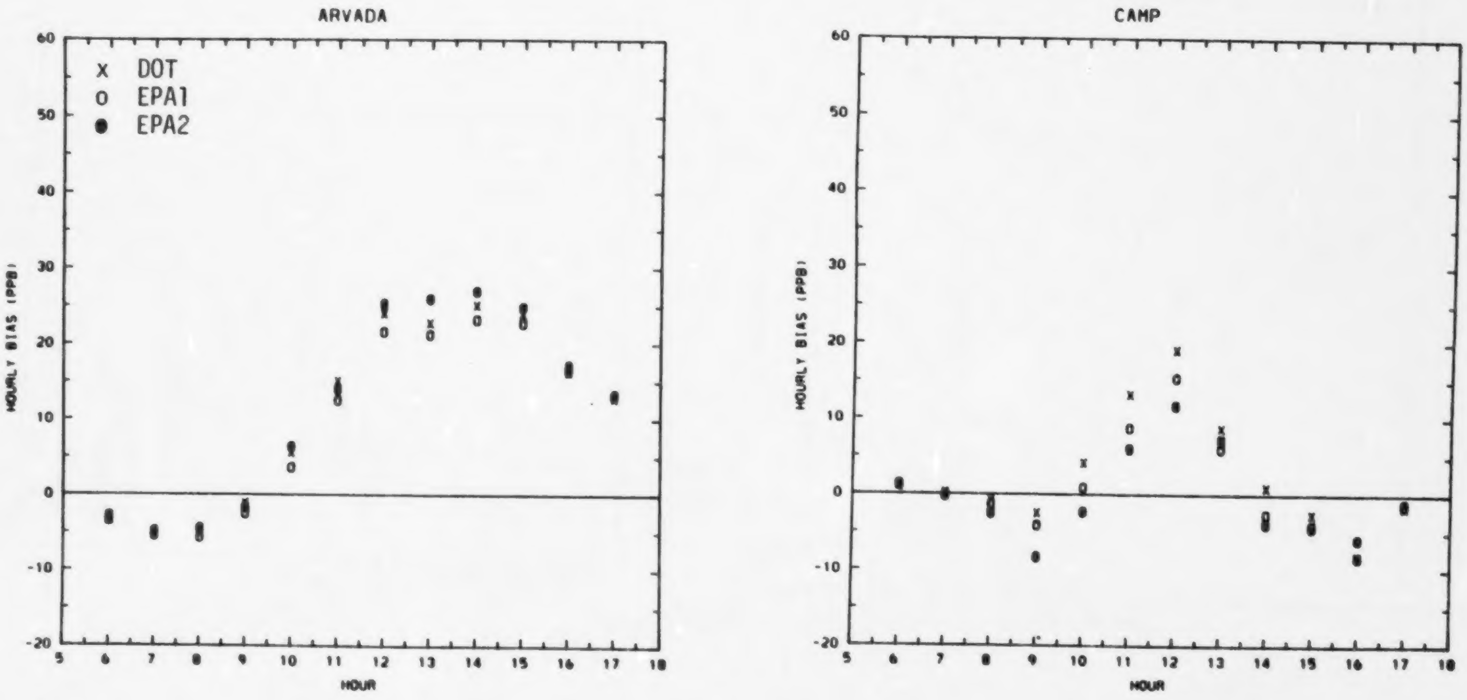


FIGURE 7 (a): Hourly Bias, Averaged Over the 11 Days, of the Three Models for the 5 Monitoring Sites.

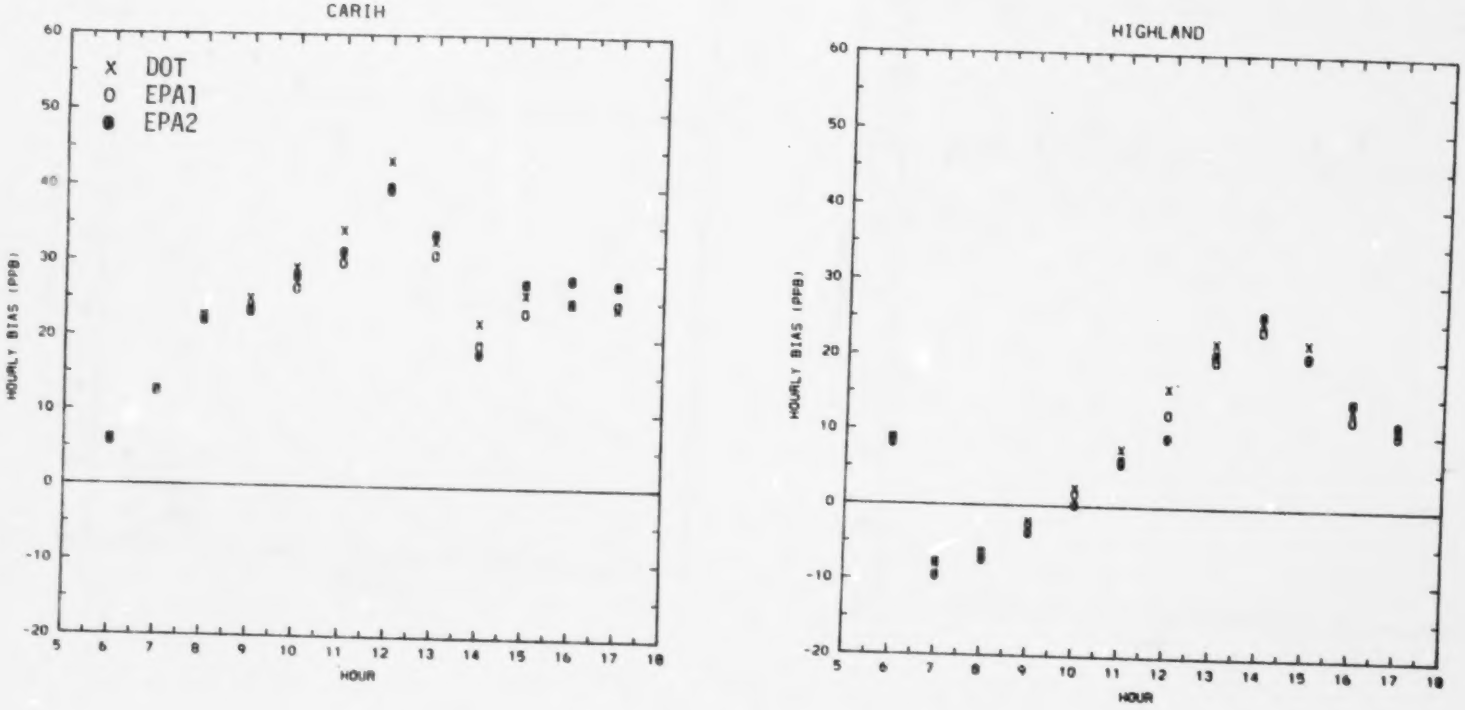


FIGURE 7 (a), (Continued)

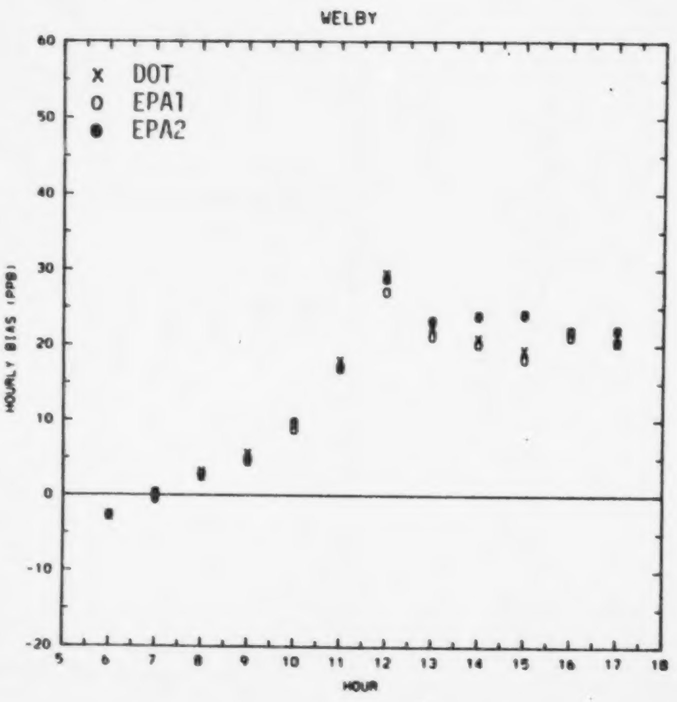


FIGURE 7 (a), Continued

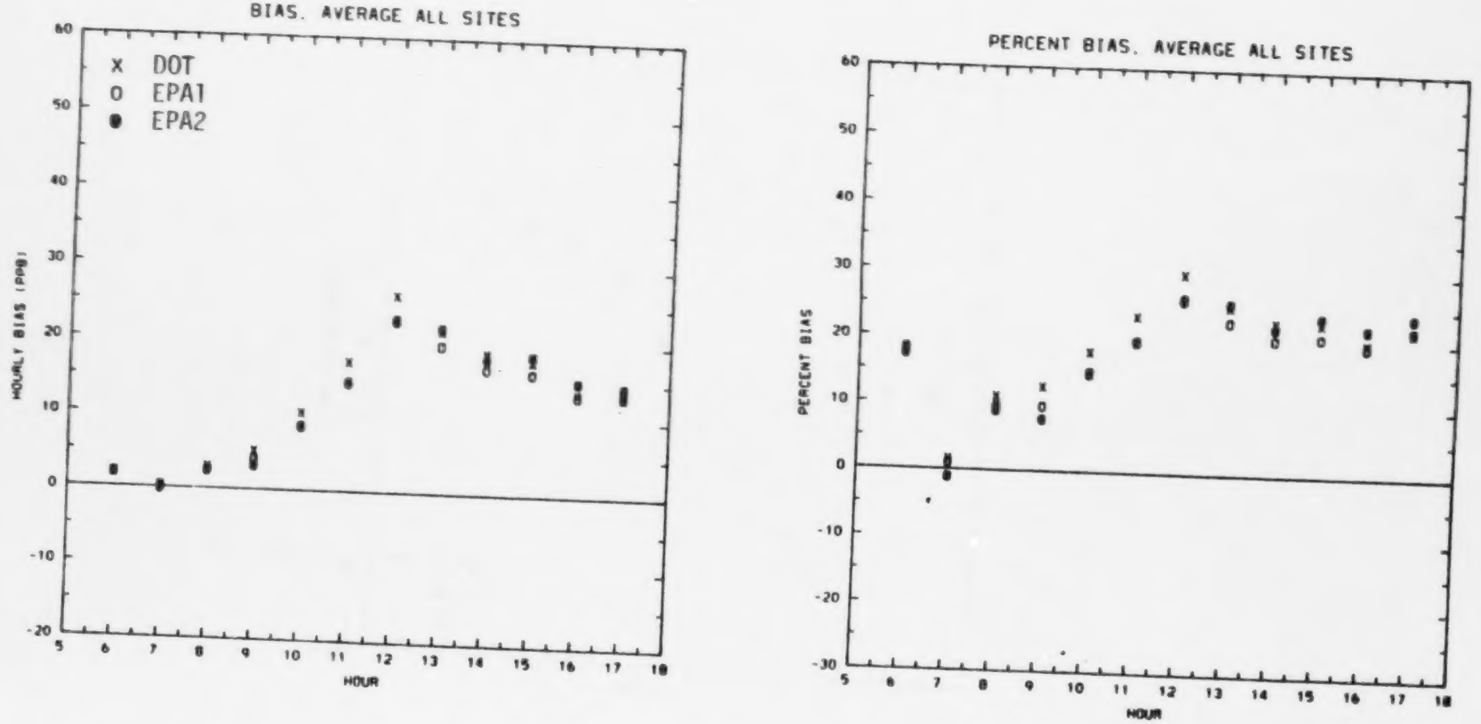


FIGURE 7 (b): Hourly Bias and % Hourly Bias Averaged Over the 11 Days and 5 Monitoring Sites.

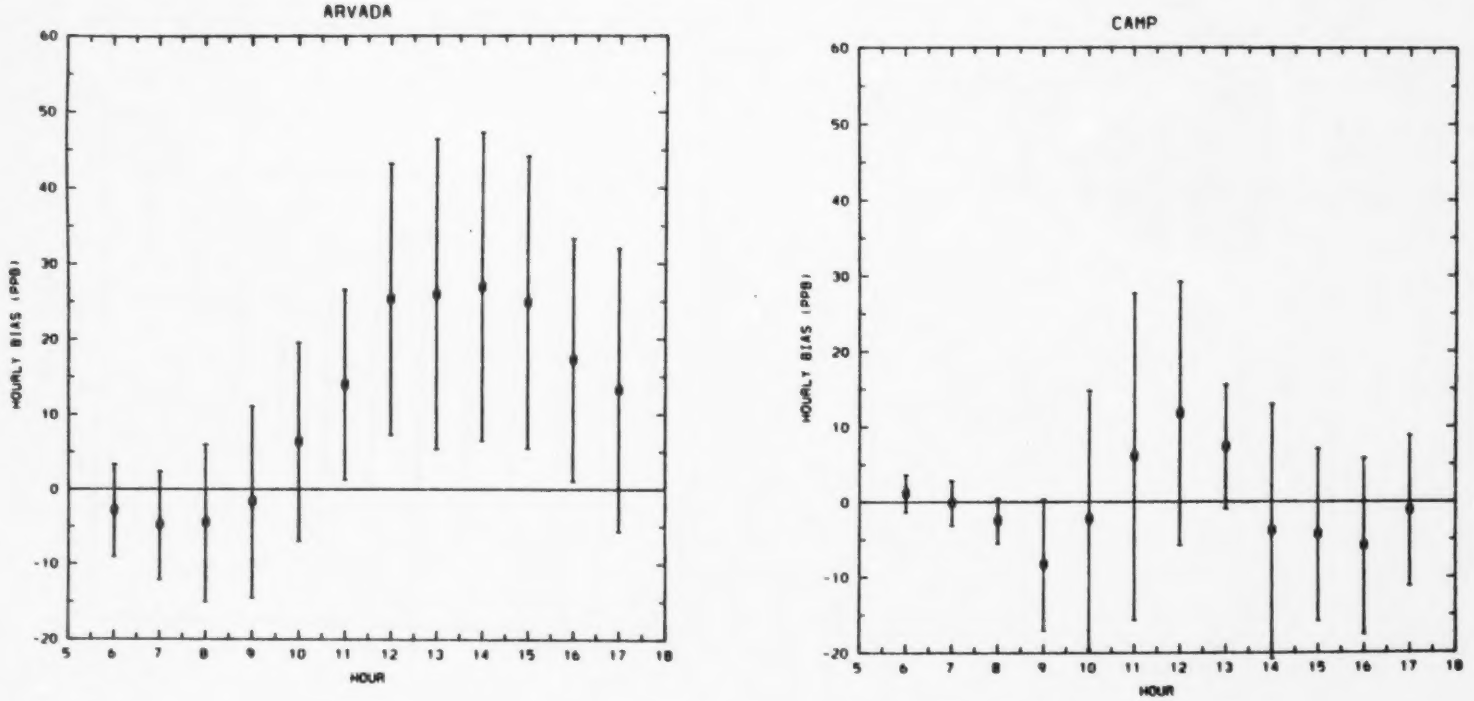


FIGURE 8: Hourly Bias for EPA2, with 95% Confidence Intervals, at Each Monitoring Site, Averaged over the 11 Days.

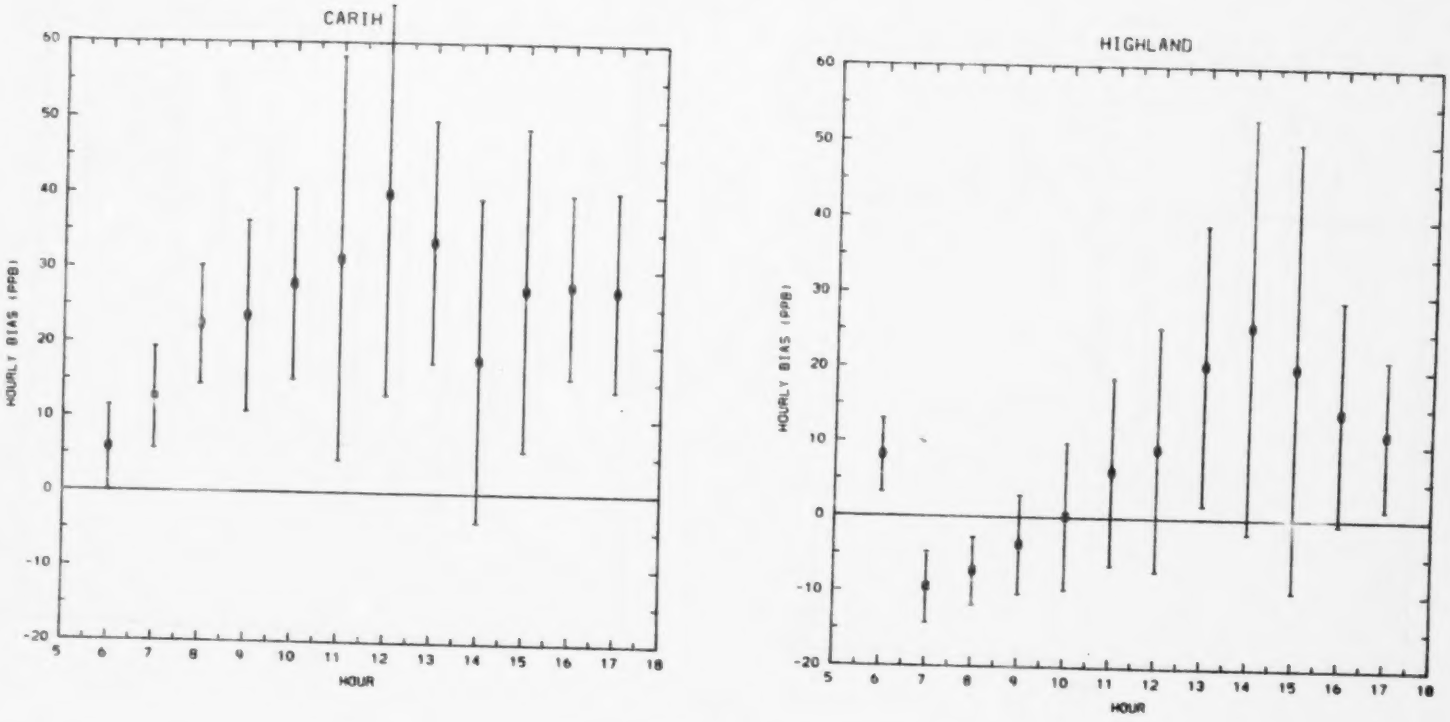


FIGURE 8, (Continued)

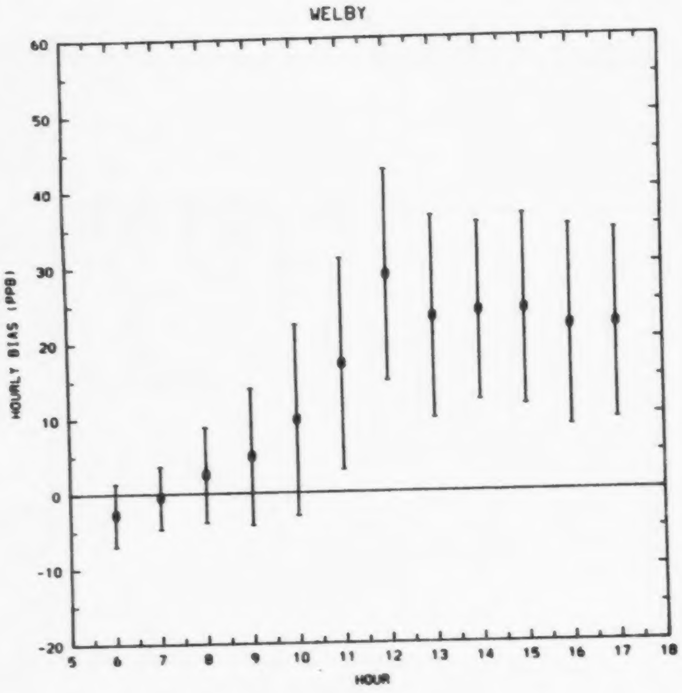


FIGURE 8, (Continued)

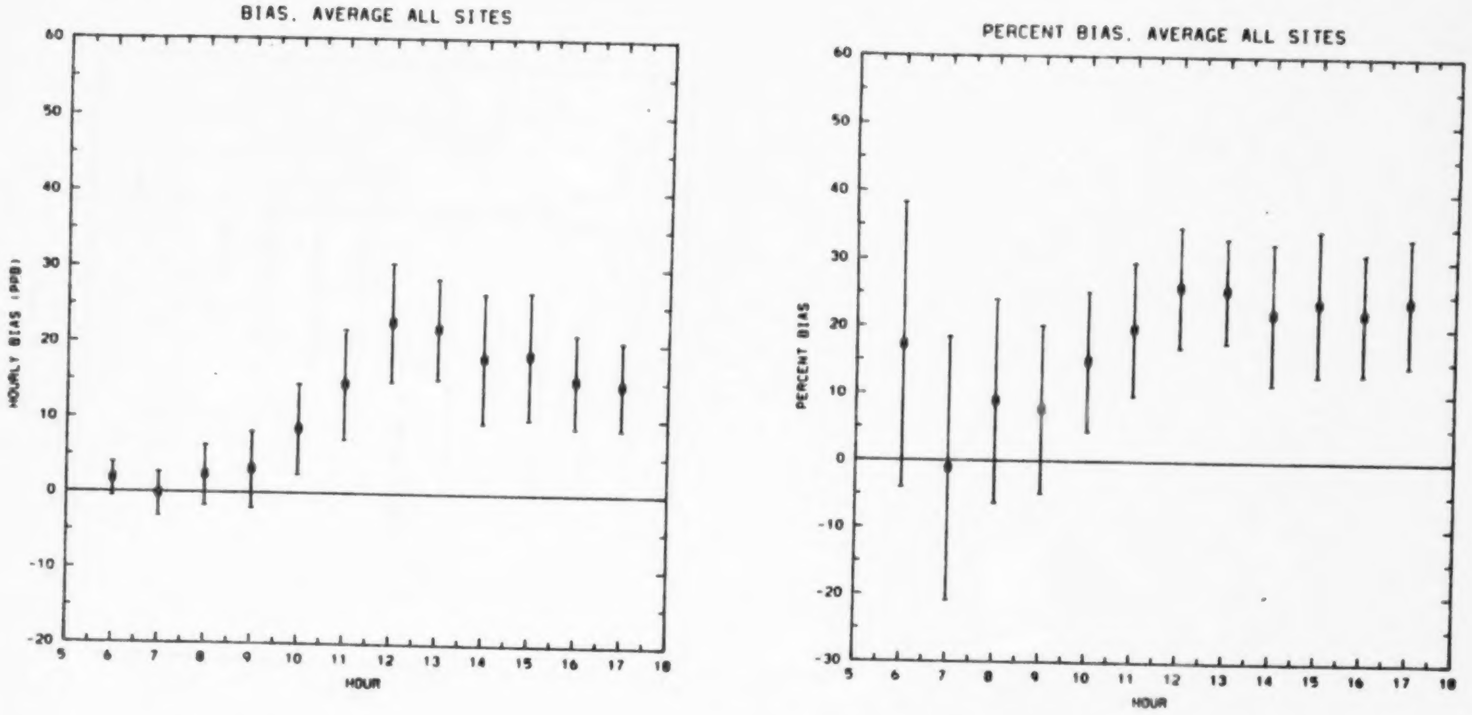


FIGURE 9: Hourly Bias and Hourly % Bias for EPA2, with 95% Confidence Intervals, Averaged over the 5 Monitoring Sites and the 11 Days.

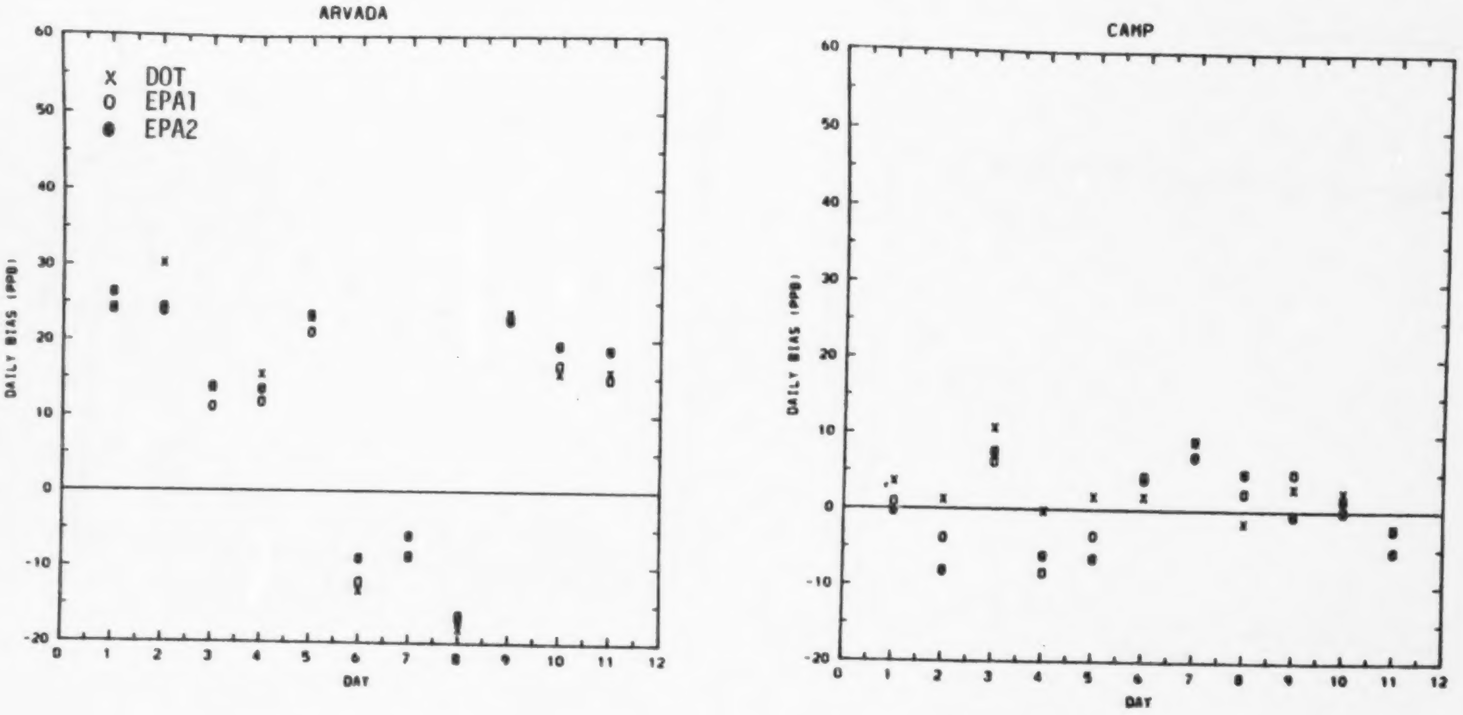


FIGURE 10: Daily Bias for the Three Models for Each of the 5 Monitoring Sites.

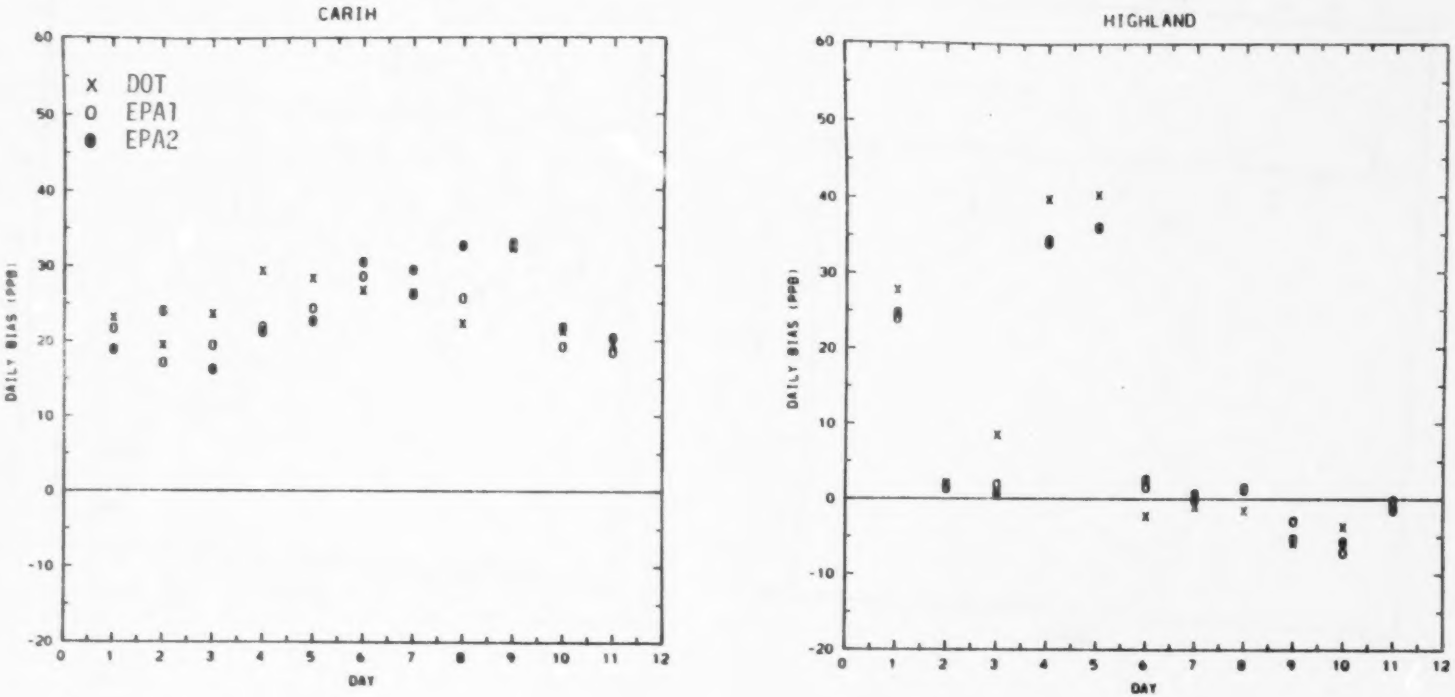


FIGURE 10, (Continued)

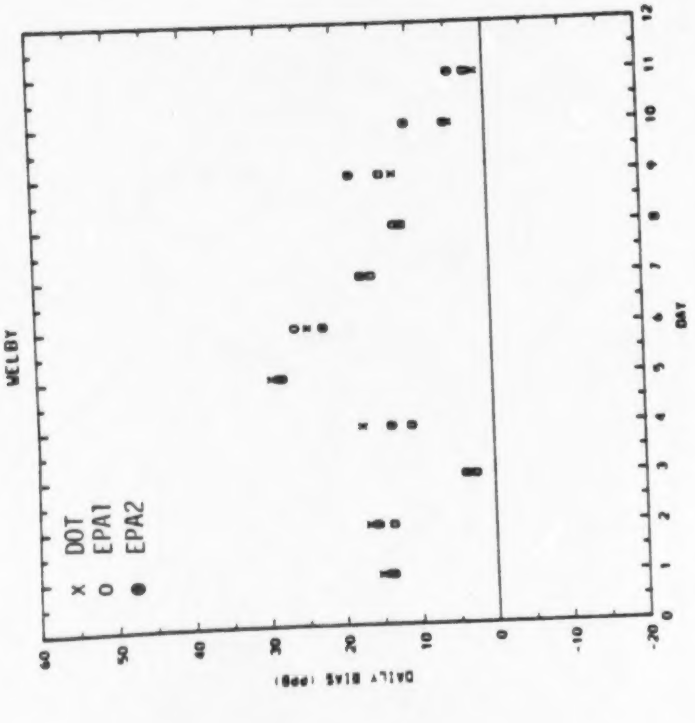


FIGURE 10, (Continued)

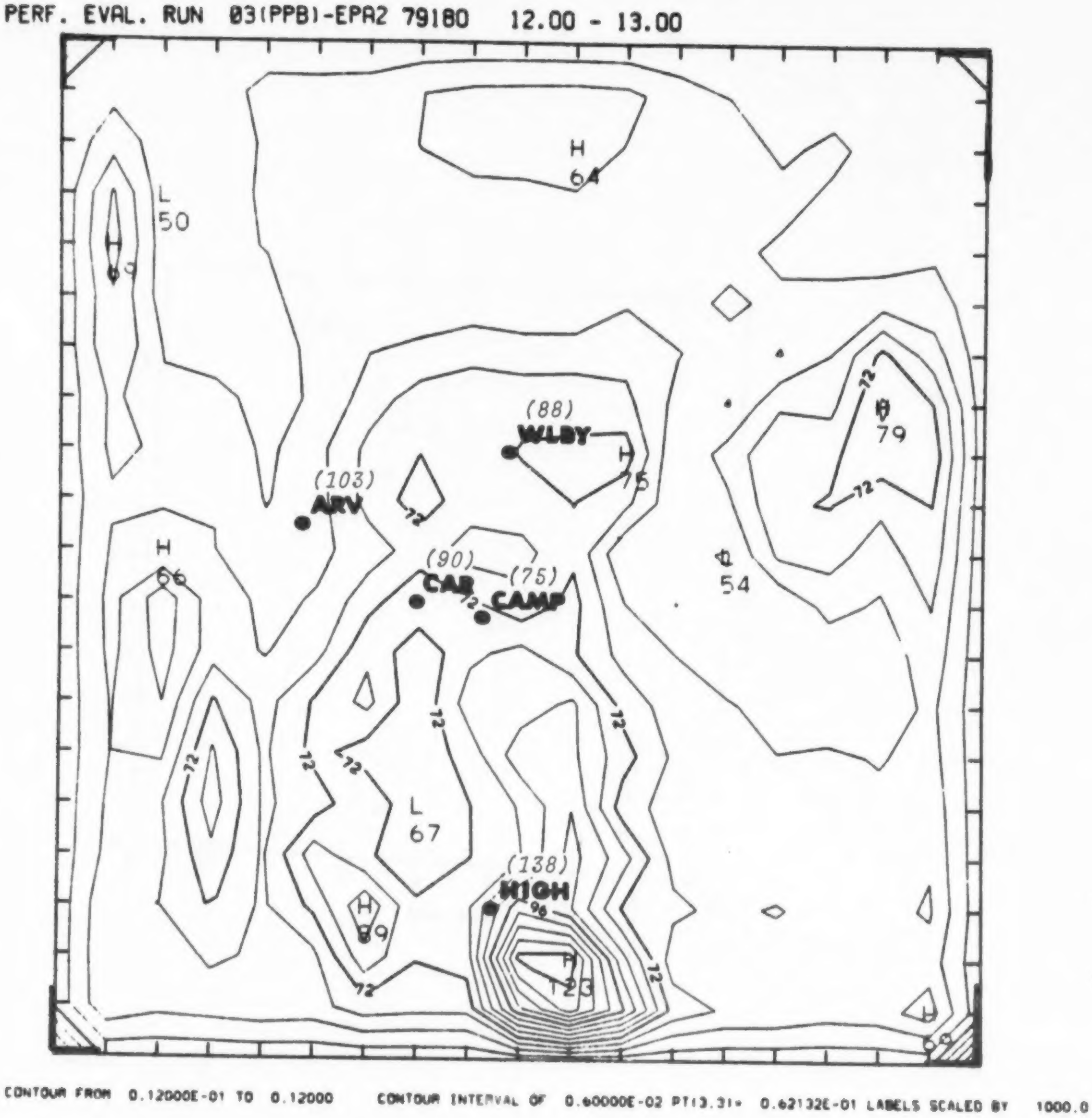


FIGURE 11: Contour Plot Showing Predicted Ozone Cloud Relative to Highland: Day 79180; Observations Are Given in Parenthesis.

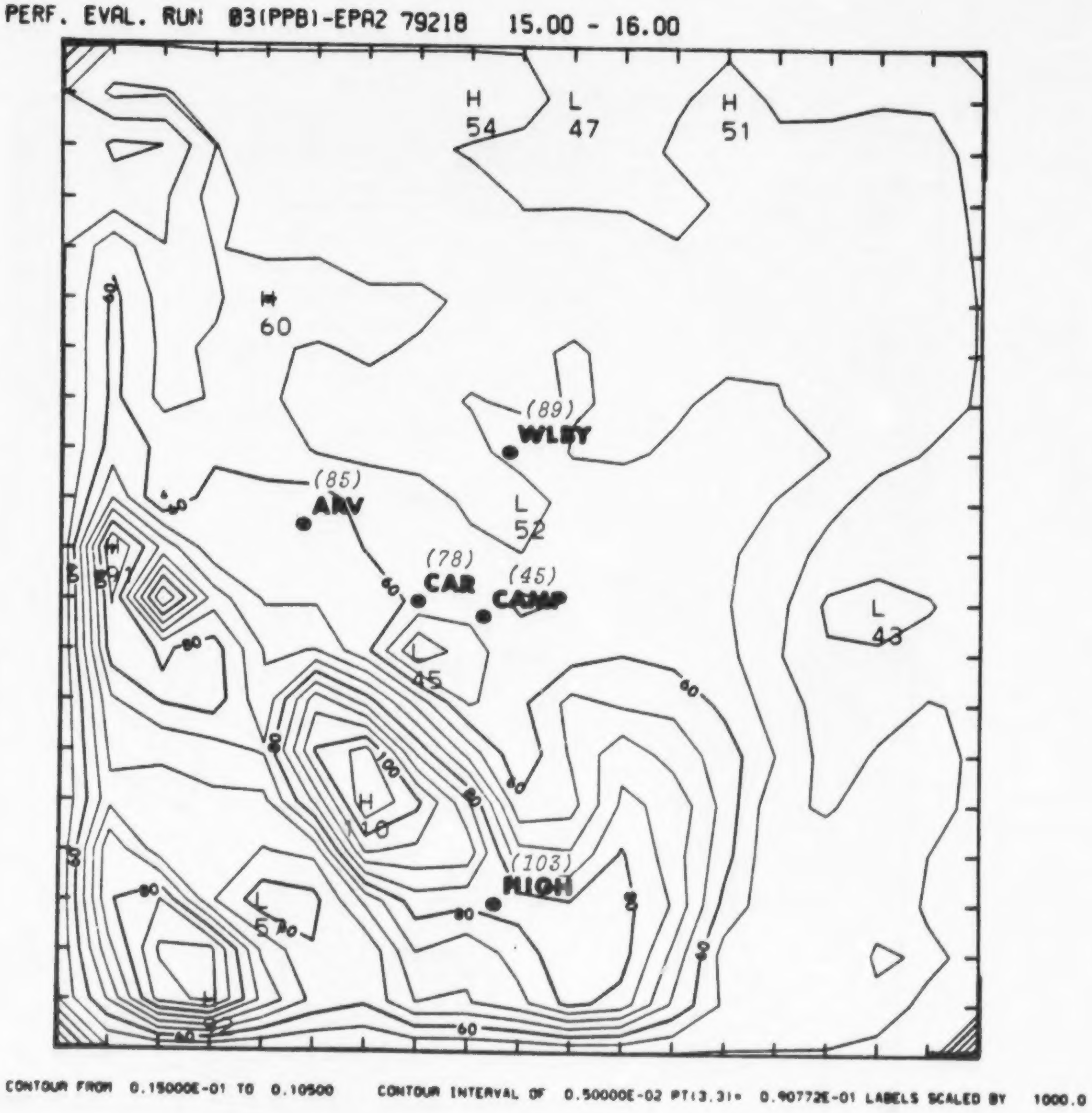


FIGURE 12: Contour Plot Showing Predicted Ozone Cloud Relative to Highland: Day 79218; Observations Are Given in Parenthesis.

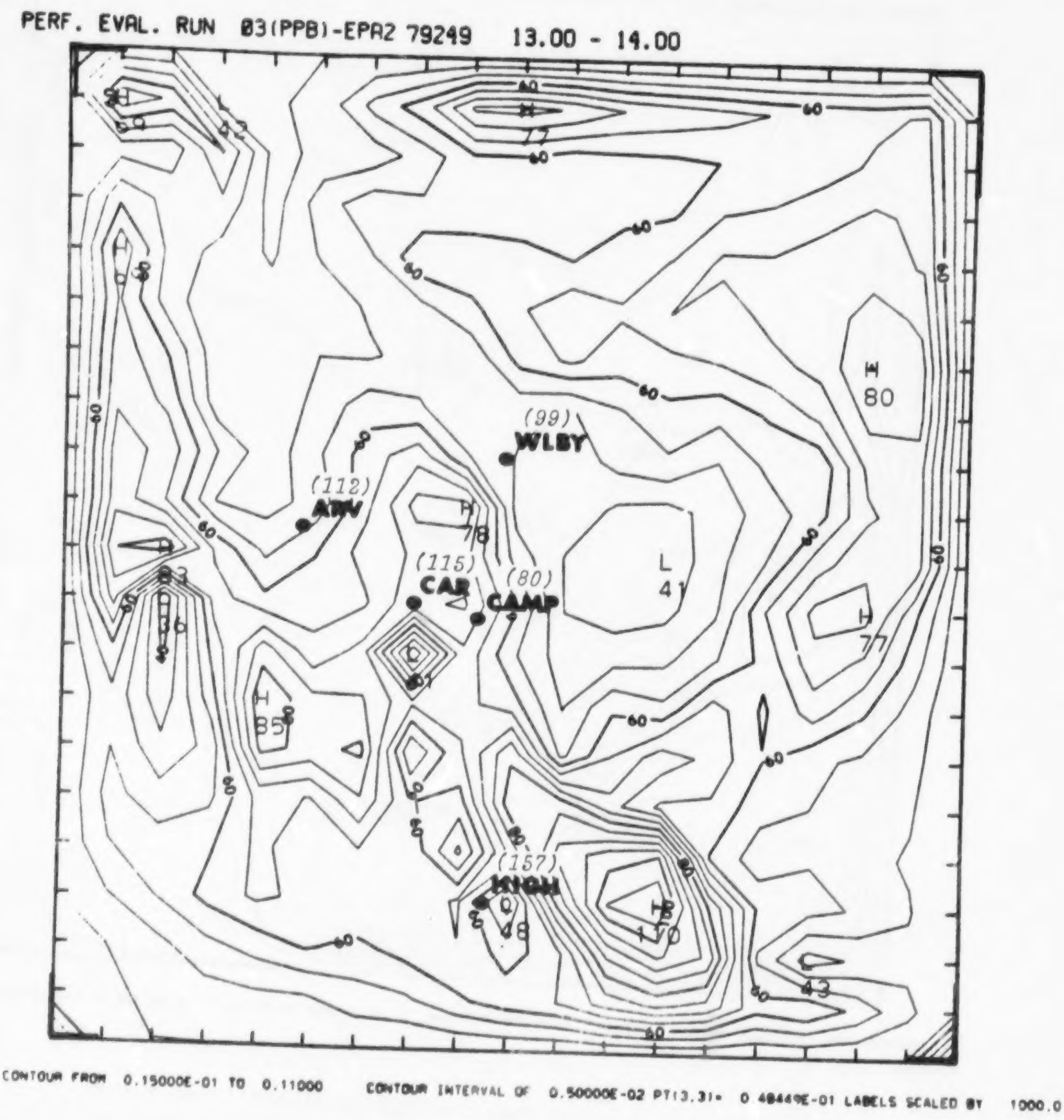

PERF. EVAL. RUN O3(PPB)-EPA2 79249 13.00 - 14.00
(99)
WLBY
(112)
ARV
(115)
CAR
(80)
CAMP
(157)
HIGH
H
80
H
77
H
78
L
41
H
85
H
110
L
43
L
48
60
CONTOUR FROM 0.15000E-01 TO 0.11000 CONTOUR INTERVAL OF 0.50000E-02 PT(3,3)= 0.48449E-01 LABELS SCALED BY 1000.0

FIGURE 13: Contour Plot Showing Predicted Ozone Cloud Relative to Highland: Day 79249; Observations Are Given in Parenthesis.

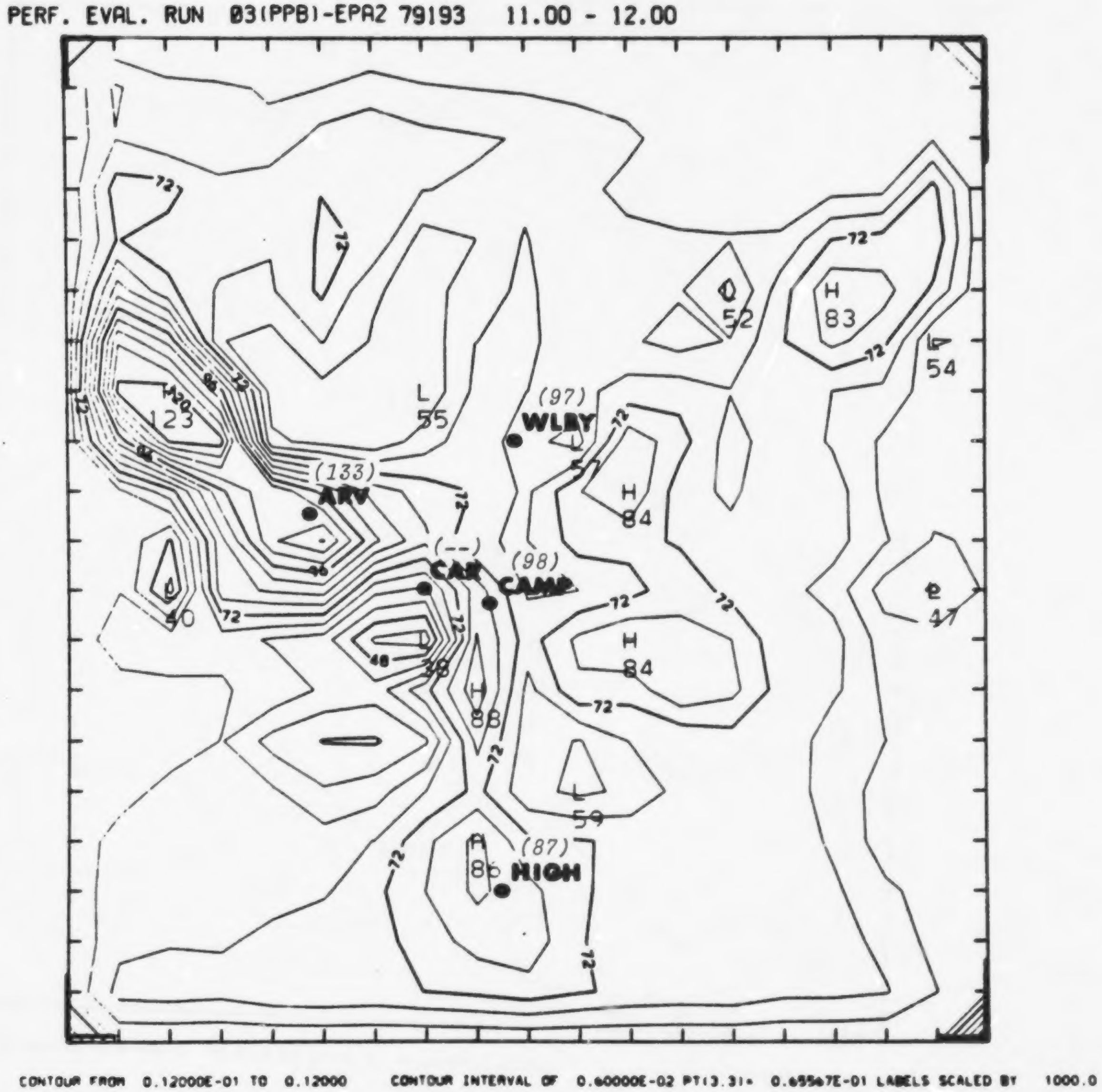

PERF. EVAL. RUN 03(PPB)-EPA2 79193 11.00 - 12.00
72
H120
123
L
55
(97)
WLBY
57
H
84
(133)
ARV
(--)
CAR
(98)
CAMP
L
40
48
L
38
H
88
H
84
52
H
83
54
47
L
59
(87)
HIGH
H
86
CONTOUR FROM 0.12000E-01 TO 0.12000 CONTOUR INTERVAL OF 0.60000E-02 PT(3,3)= 0.65567E-01 LABELS SCALED BY 1000.0

FIGURE 14: Contour Plot Showing Predicted Ozone Cloud Relative to Arvada: Day 79193; Observations Are Given in Parenthesis.

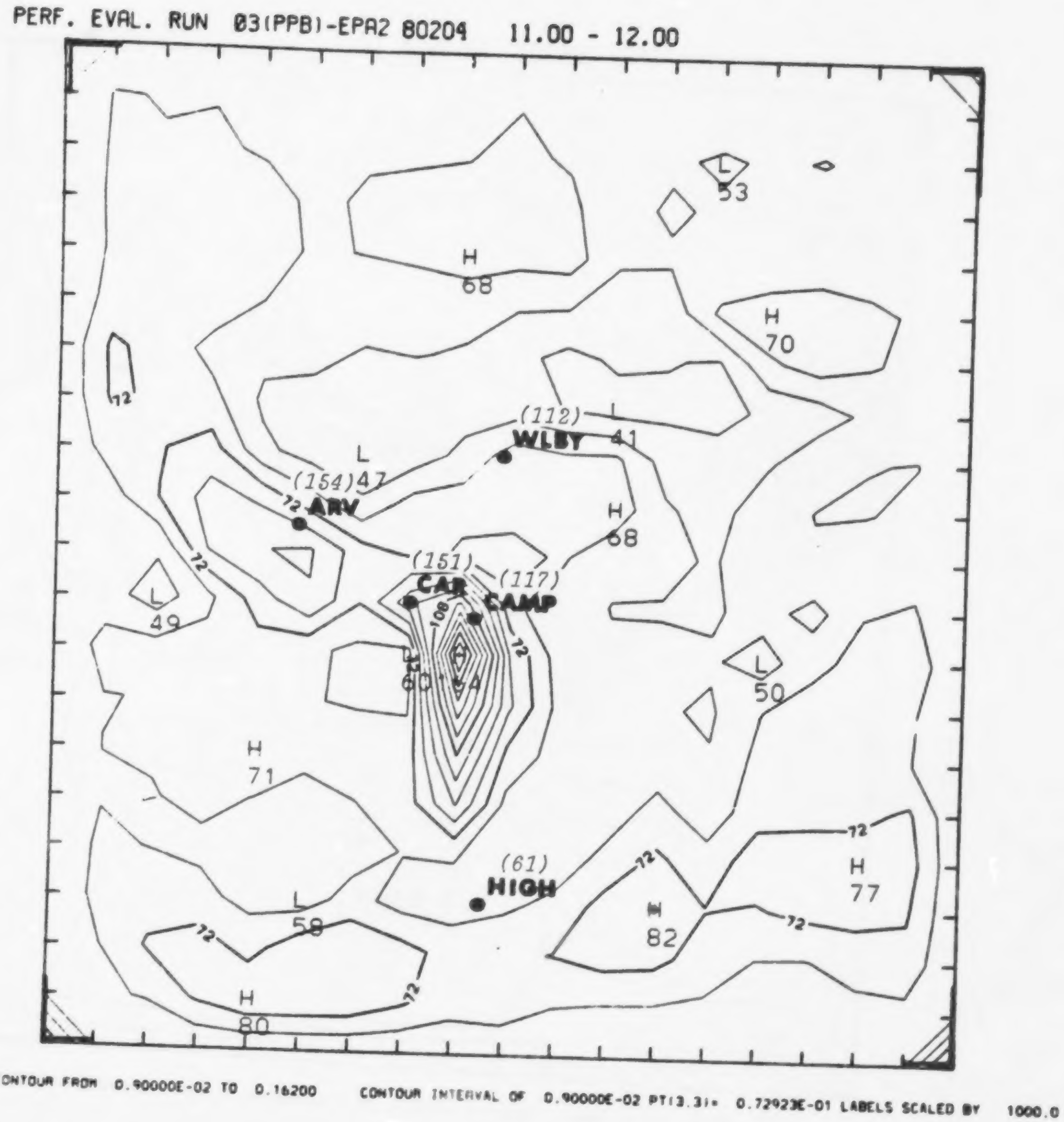
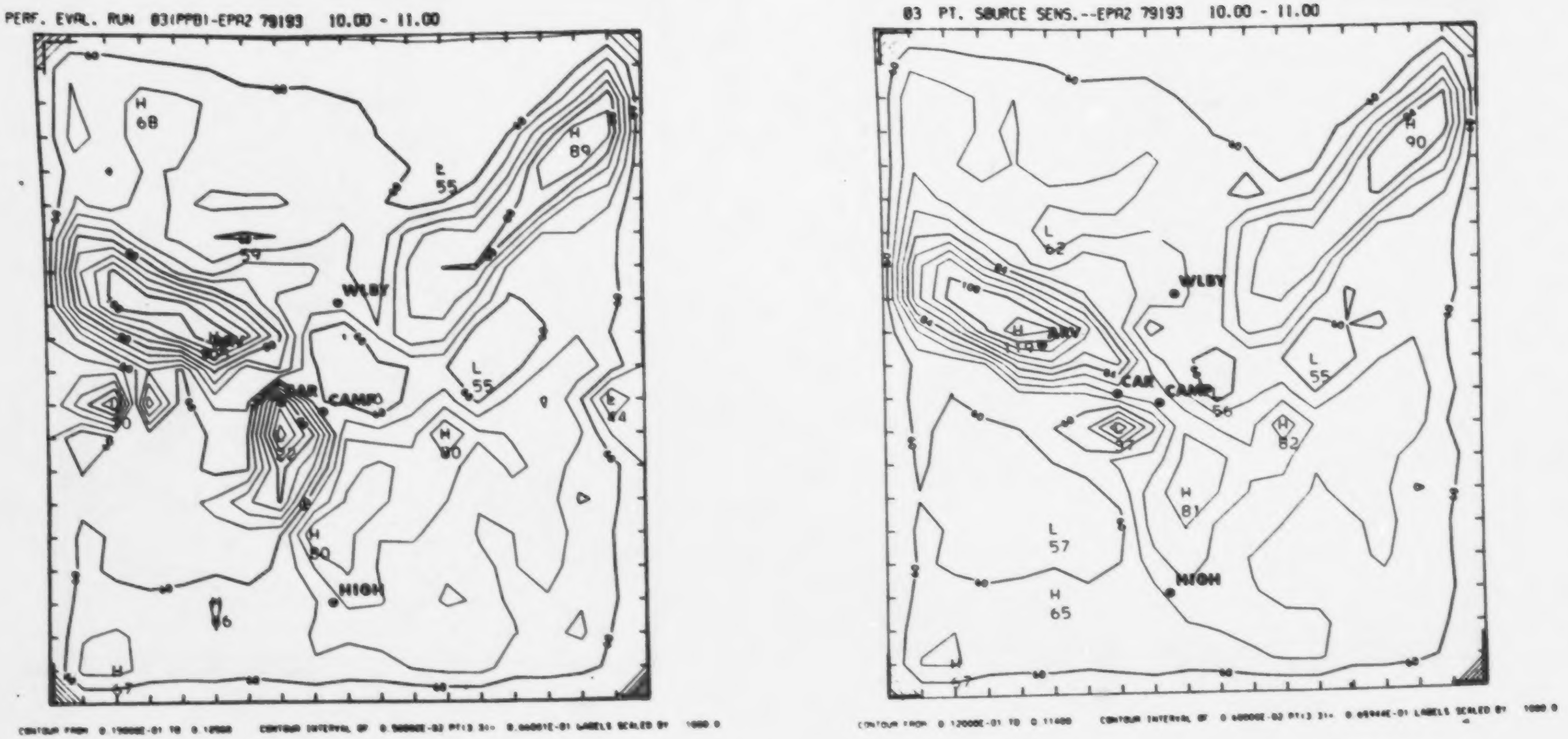

PERF. EVAL. RUN 03(PPB)-EPA2 80204 11.00 - 12.00
L
53
H
68
H
70
(112)
WLBY
L
41
L
47
(154)
ARV
72
H
68
(151)
CAR
(117)
CAMP
108
72
L
49
L
50
H
71
(61)
HIGH
72
H
77
H
82
L
58
H
80
CONTOUR FROM 0.90000E-02 TO 0.16200 CONTOUR INTERVAL OF 0.90000E-02 PT(3.3)= 0.72923E-01 LABELS SCALED BY 1000.0

FIGURE 15: Contour Plot Showing Predicted Ozone Cloud Relative to Arvada: Day 80204; Observations Are Given in Parenthesis.



PERF. EVAL. RUN 03(PPB)-EPA2 79193 10.00 - 11.00
H 68
L 55
H 89
59
WLBY
L 55
CAR
CAMP
H 80
H 80
HIGH
03 PT. SOURCE SENS.--EPA2 79193 10.00 - 11.00
H 90
L 62
WLBY
ARV
L 55
CAR
CAMP
56
H 82
37
H 81
L 57
HIGH
H 65
CONTOUR FROM 0.12000E-01 TO 0.11400 CONTOUR INTERVAL OF 0.60000E-02 PT(3 3)= 0.65944E-01 LABELS SCALED BY 1000.0

FIGURE 16: Comparison of Simulations with Point Sources (left) and Without Point Sources (right) Day 79193, Hour 11.

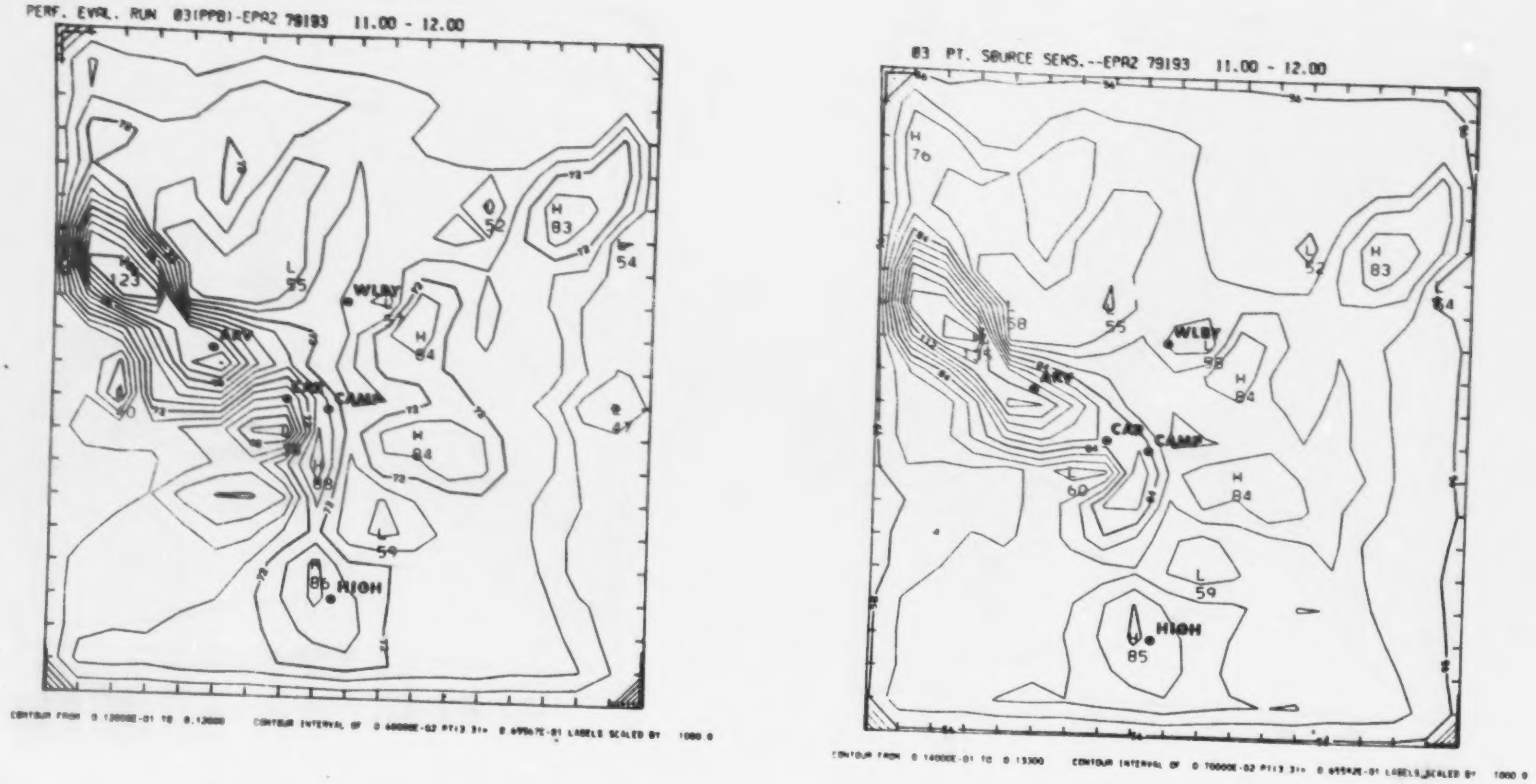

PERF. EVAL. RUN O3(PPB)-EPA2 79193 11.00 - 12.00
H
123
L
55
52
H
83
54
WLBY
H
84
ARY
CAR
CAMA
H
84
47
H
88
L
59
86
HIGH
CONTOUR FROM 0.12000E-01 TO 0.12000 CONTOUR INTERVAL OF 0.40000E-02 LABELS SCALED BY 1000.0
O3 PT. SOURCE SENS.--EPA2 79193 11.00 - 12.00
H
76
L
58
134
55
L
52
H
83
L
54
WLBY
L
98
H
84
ARY
CAR
CAMP
L
60
H
84
L
59
HIGH
85
CONTOUR FROM 0.14000E-01 TO 0.13300 CONTOUR INTERVAL OF 0.70000E-02 LABELS SCALED BY 1000.0

FIGURE 17: Comparison of Simulation with Point Sources (left) and Without Point Sources (right) Day 79193, Hour 12.

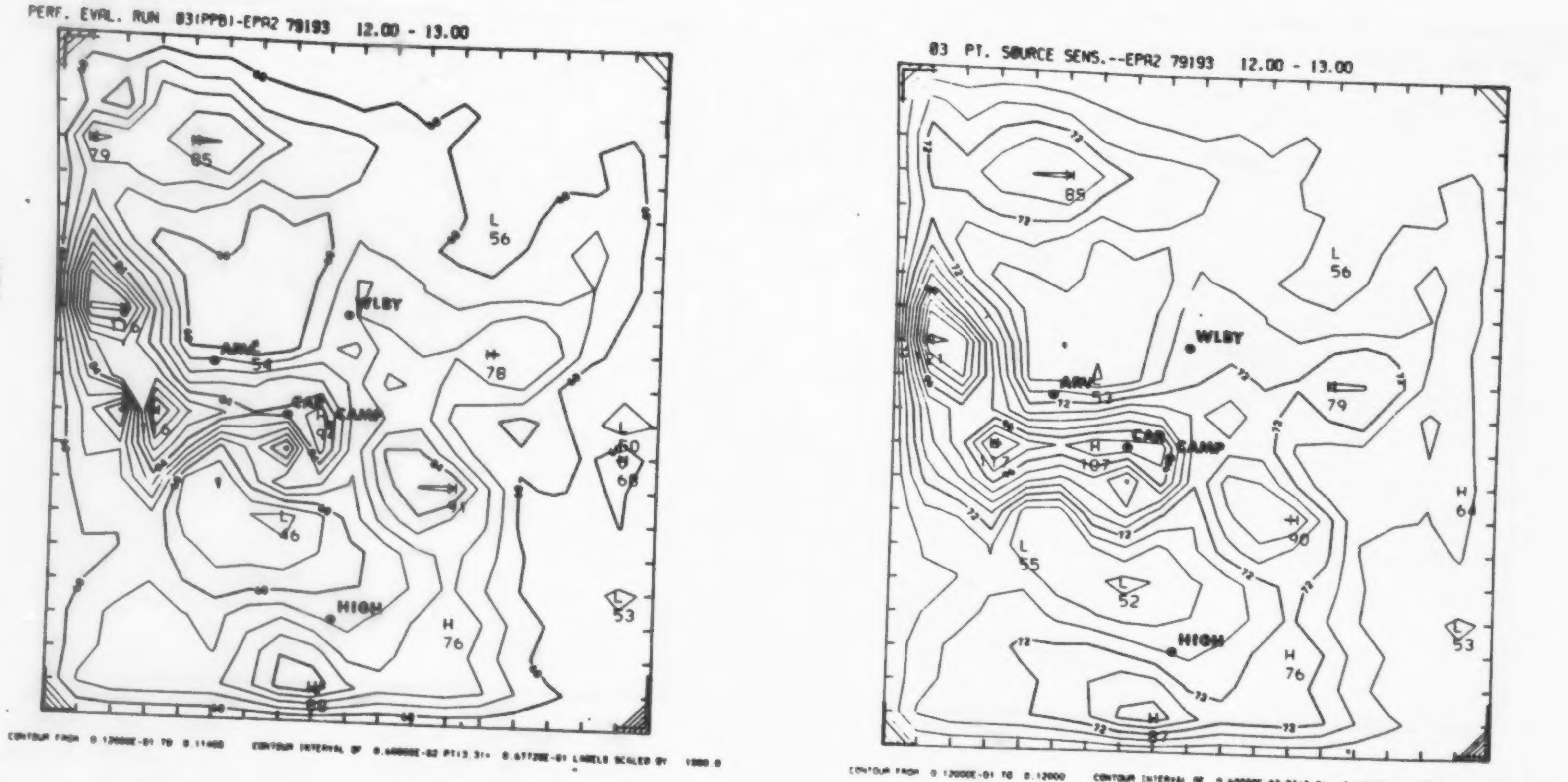
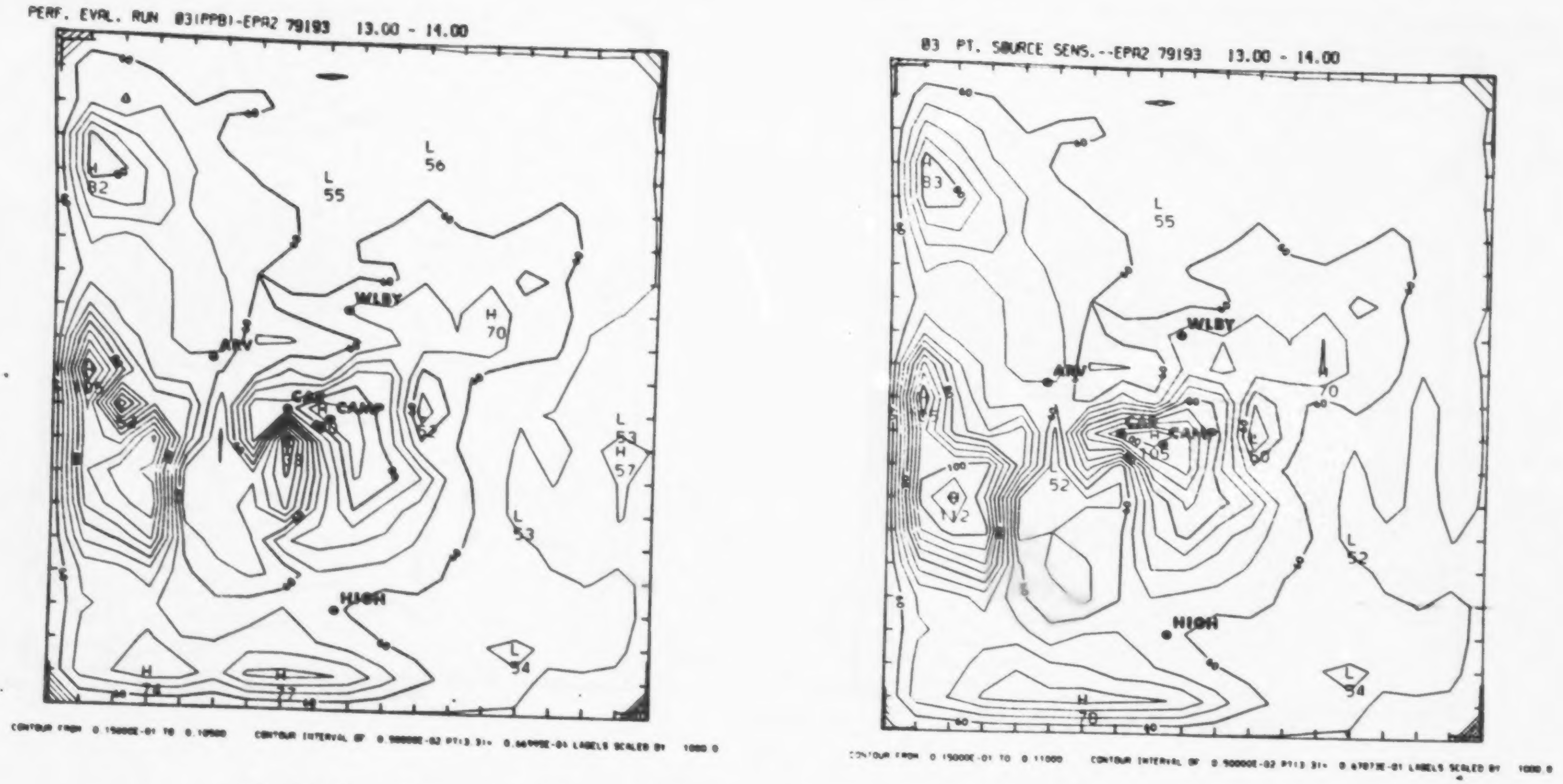

PERF. EVAL. RUN 03(PPB)-EPA2 79193 12.00 - 13.00
79
85
L
56
WLBY
ARVC
54
78
CAR
CAMP
60
68
46
HIGH
H
76
53
CONTOUR FROM 0.12000E-01 TO 0.11400 CONTOUR INTERVAL OF 0.60000E-02 PT(3,3)= 0.67728E-01 LABELS SCALED BY 1000.0
03 PT. SOURCE SENS.--EPA2 79193 12.00 - 13.00
72
85
L
56
WLBY
ARVC
53
79
H
107
CAR
CAMP
H
64
90
L
55
52
HIGH
H
76
53
87
CONTOUR FROM 0.12000E-01 TO 0.12000 CONTOUR INTERVAL OF 0.60000E-02 PT(3,3)= 0.67674E-01 LABELS SCALED BY 1000.0

FIGURE 18: Comparison of Simulation with Point Sources (left) and Without Point Sources (right) Day 79193, Hour 13.



PERF. EVAL. RUN O3(PPB)-EPA2 79193 13.00 - 14.00
H 82
L 55
L 56
WLBY
H 70
ARV
CAR
CAMP
52
L 53
H 57
L 53
HIGH
L 54
H 78
H 77
CONTOUR FROM 0.15000E-01 TO 0.10500 CONTOUR INTERVAL OF 0.90000E-02 PT(3,3)= 0.66995E-01 LABELS SCALED BY 1000.0
O3 PT. SOURCE SENS.--EPA2 79193 13.00 - 14.00
83
L 55
WLBY
70
ARV
CAR
CAMP
H 105
60
100
112
52
L 52
HIGH
L 54
H 70
CONTOUR FROM 0.15000E-01 TO 0.11000 CONTOUR INTERVAL OF 0.90000E-02 PT(3,3)= 0.67073E-01 LABELS SCALED BY 1000.0

FIGURE 19: Comparison of Simulation with Point Sources (left) and Without Point Sources (right) Day 79193, Hour 14.

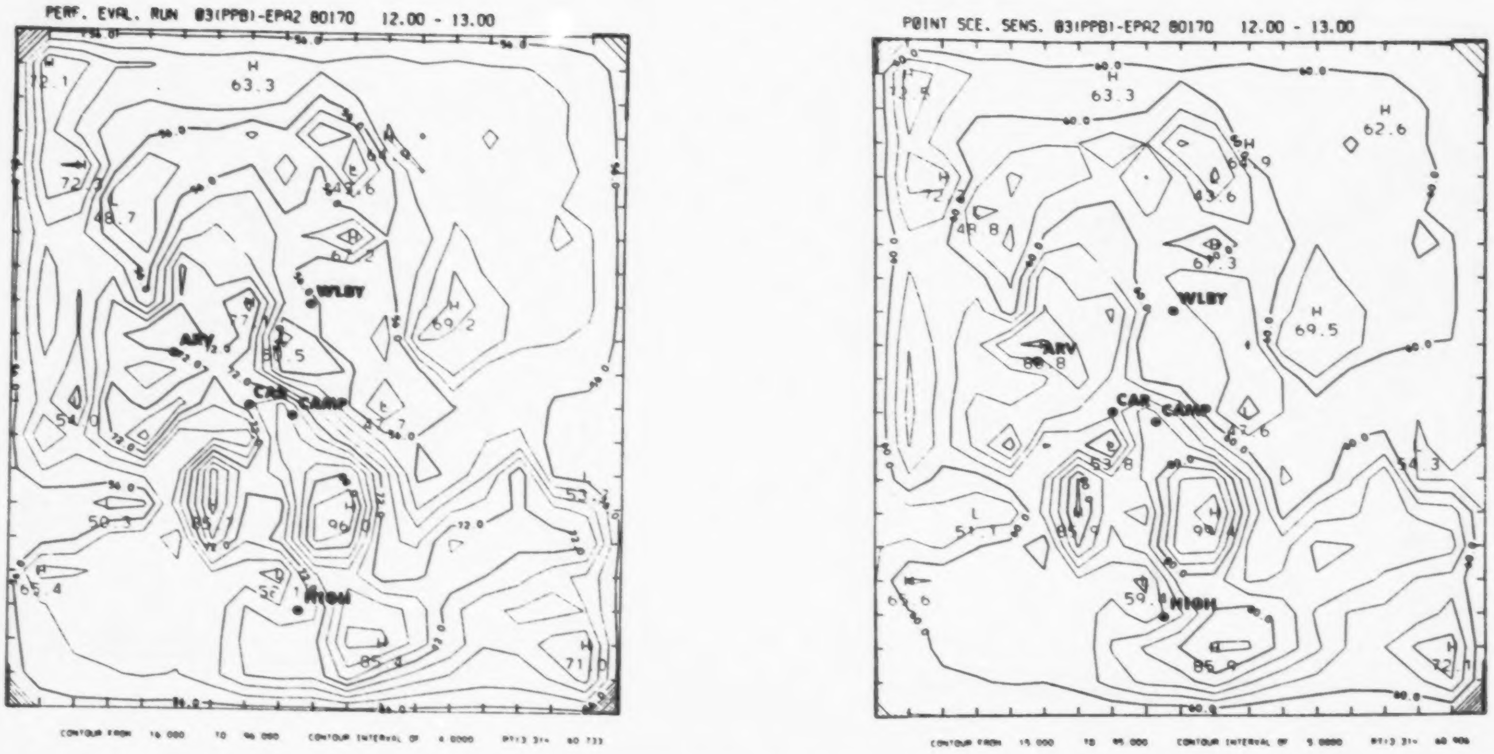

PERF. EVAL. RUN 03(PPB)-EPA2 80170 12.00 - 13.00
H 63.3
72.1
72.3
48.7
49.6
67.2
WLBY
69.2
ARV
77.7
80.5
CAR
CAMP
54.0
52.7
50.3
85.7
90.0
56.1
HIGH
65.4
85.4
71.0
CONTOUR FROM 16.000 TO 96.000 CONTOUR INTERVAL OF 4.0000 PT(3,3)= 60.733
POINT SCE. SENS. 03(PPB)-EPA2 80170 12.00 - 13.00
H 63.3
H 62.6
72.9
68.9
72.2
48.8
43.6
67.3
WLBY
H 69.5
ARV
80.8
CAR
CAMP
72.6
63.8
54.3
51.1
85.9
90.4
59.8
HIGH
65.6
85.9
72.1
CONTOUR FROM 15.000 TO 95.000 CONTOUR INTERVAL OF 5.0000 PT(3,3)= 60.906

FIGURE 20: Comparison of Simulations with Point Sources (left) and Without Point Sources (right) Day 80170, Hour 13.

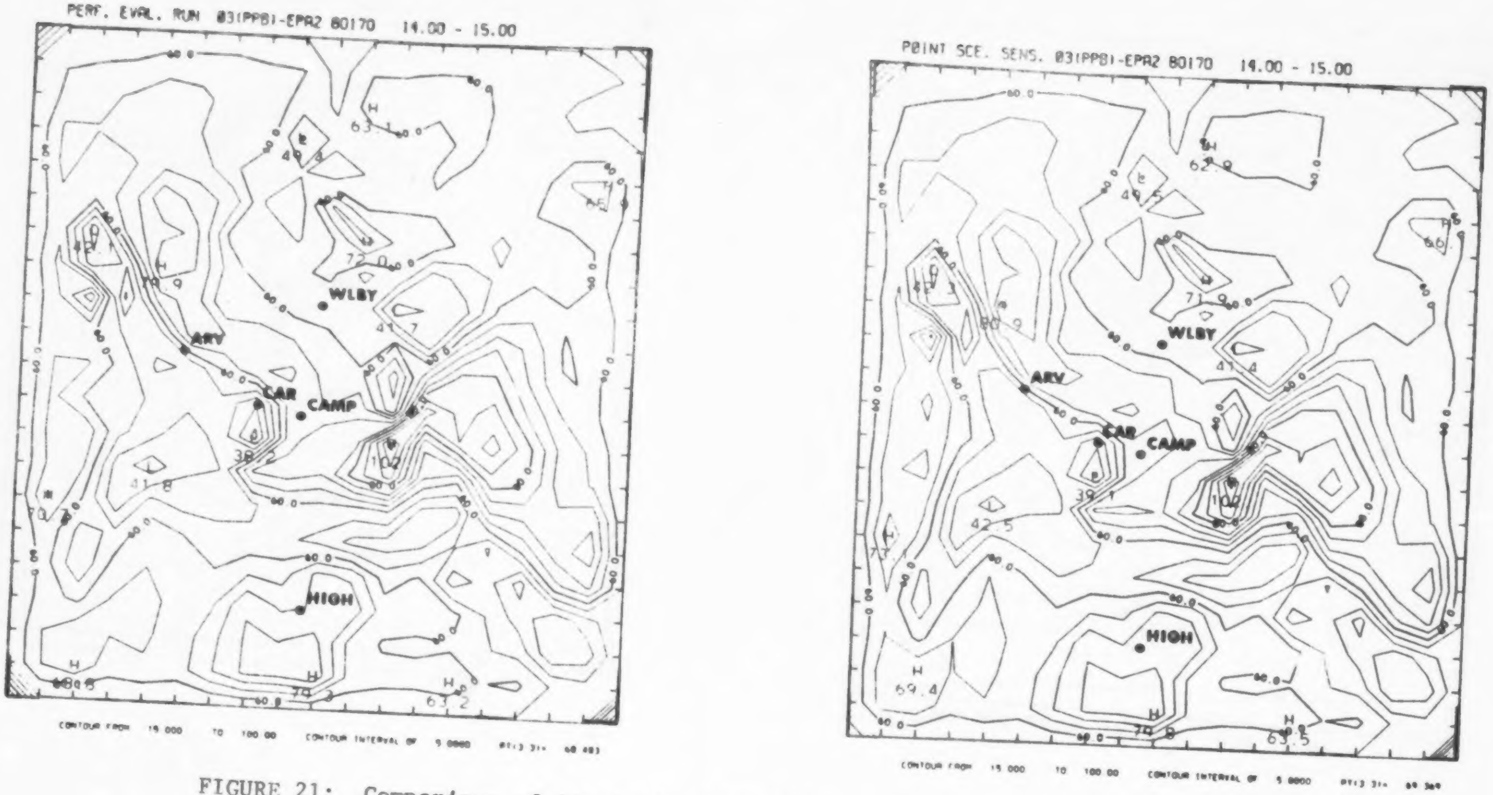

PERF. EVAL. RUN Ø3(PPB)-EPA2 80170 14.00 - 15.00
WLBY
ARV
CAR
CAMP
HIGH
H
63.1
49.4
72.0
79.9
41.7
38.2
102
41.8
70.2
60.0
63.2
CONTOUR FROM 15.000 TO 100.00 CONTOUR INTERVAL OF 5.0000
PØINT SCE. SENS. Ø3(PPB)-EPA2 80170 14.00 - 15.00
WLBY
ARV
CAR
CAMP
HIGH
62.9
49.5
66.
71.9
80.9
41.4
39.1
102
42.5
69.4
70.8
63.5
CONTOUR FROM 15.000 TO 100.00 CONTOUR INTERVAL OF 5.0000

FIGURE 21: Comparison of Simulation with Point Sources (left) and Without Point Sources (right) Day 80170, Hour 15.

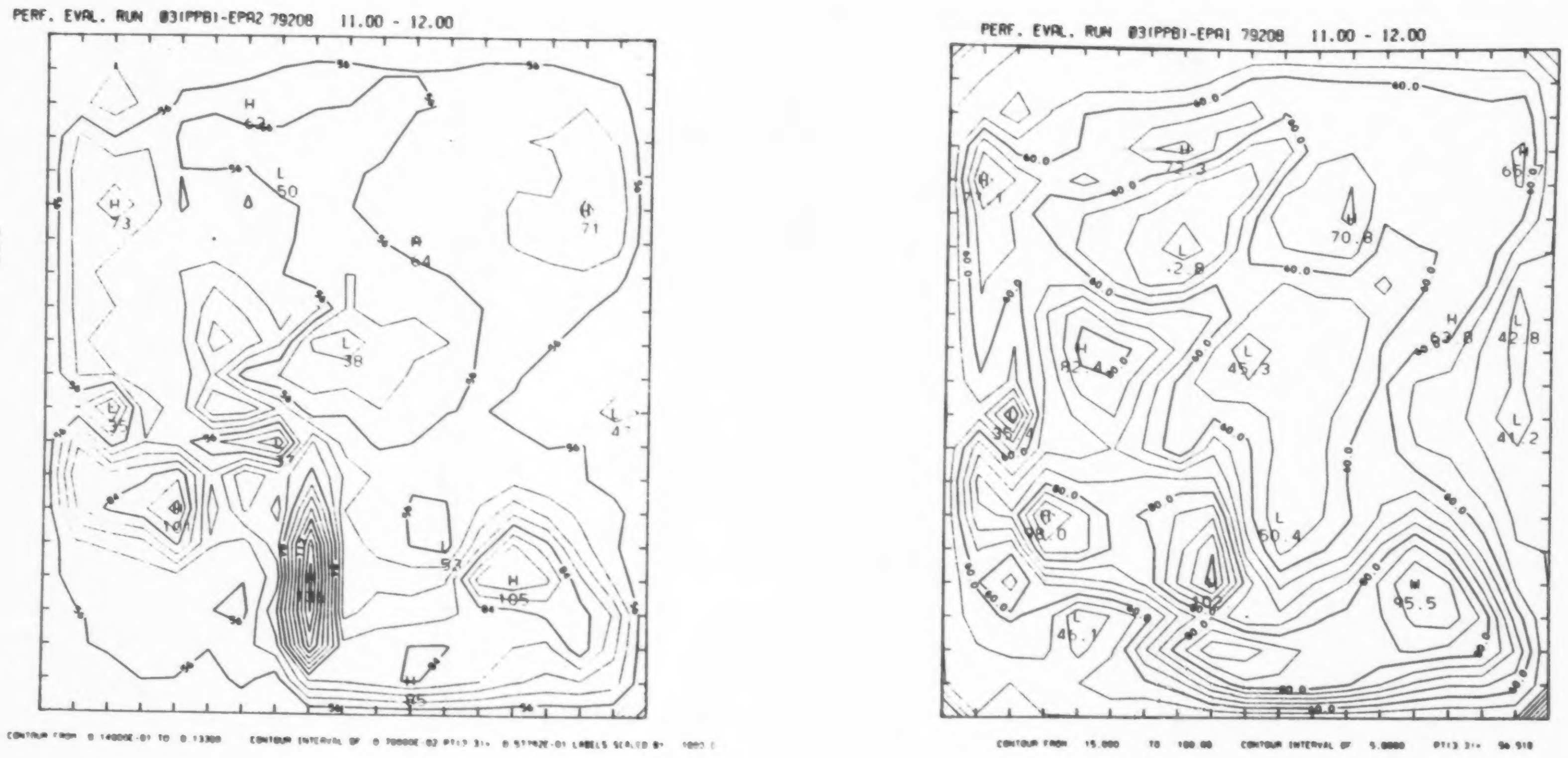

PERF. EVAL. RUN O3(PPB)-EPA2 79208 11.00 - 12.00
CONTOUR FROM 0.14000E-01 TO 0.13300 CONTOUR INTERVAL OF 0.70000E-02 PT(3,3)= 0.57792E-01 LABELS SCALED BY 1000.0
PERF. EVAL. RUN O3(PPB)-EPA1 79208 11.00 - 12.00
CONTOUR FROM 15.000 TO 100.00 CONTOUR INTERVAL OF 5.0000 PT(3,3)= 56.510

FIGURE 22: Comparison of the Simulations of EPA2 (left) and EPA1 (right) Including Point Sources with Simulation of EPA2 Excluding Point Sources (overleaf) Day 79208, Hour 12

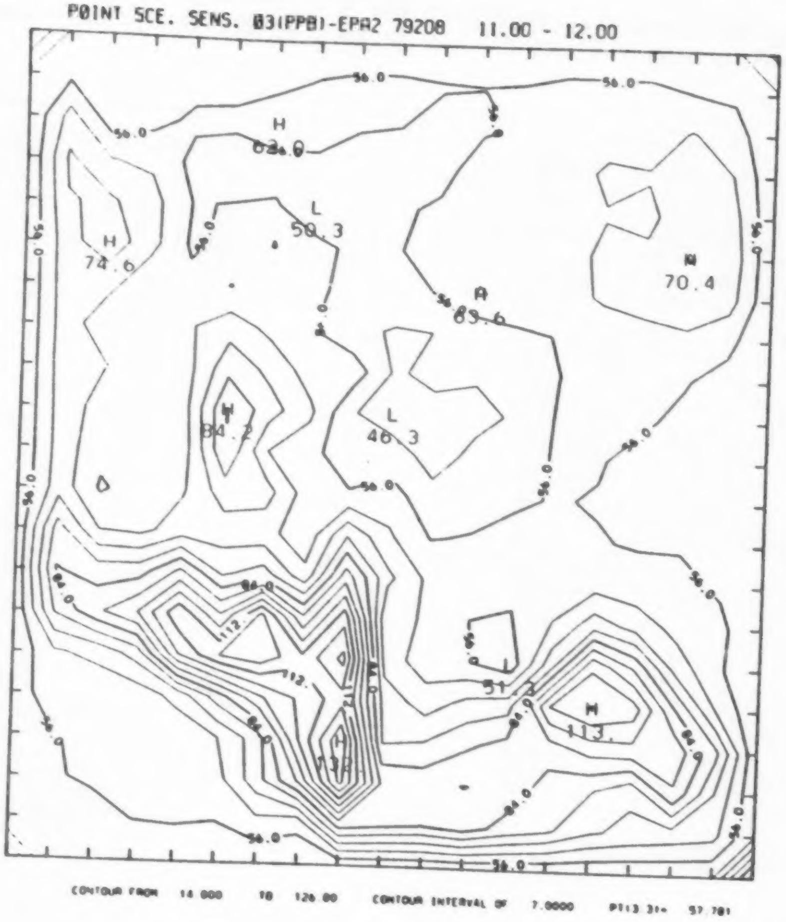

P0INT SCE. SENS. 03(PPB)-EPA2 79208 11.00 - 12.00
H
63.0
L
50.3
H
74.6
A
63.6
H
70.4
H
84.2
L
46.3
L
51.3
H
113.
H
132.
56.0
0.95
84.0
112.
CONTOUR FROM 14.000 TO 126.00 CONTOUR INTERVAL OF 7.0000 PT(3,3)= 57.701

FIGURE 22, (Continued)

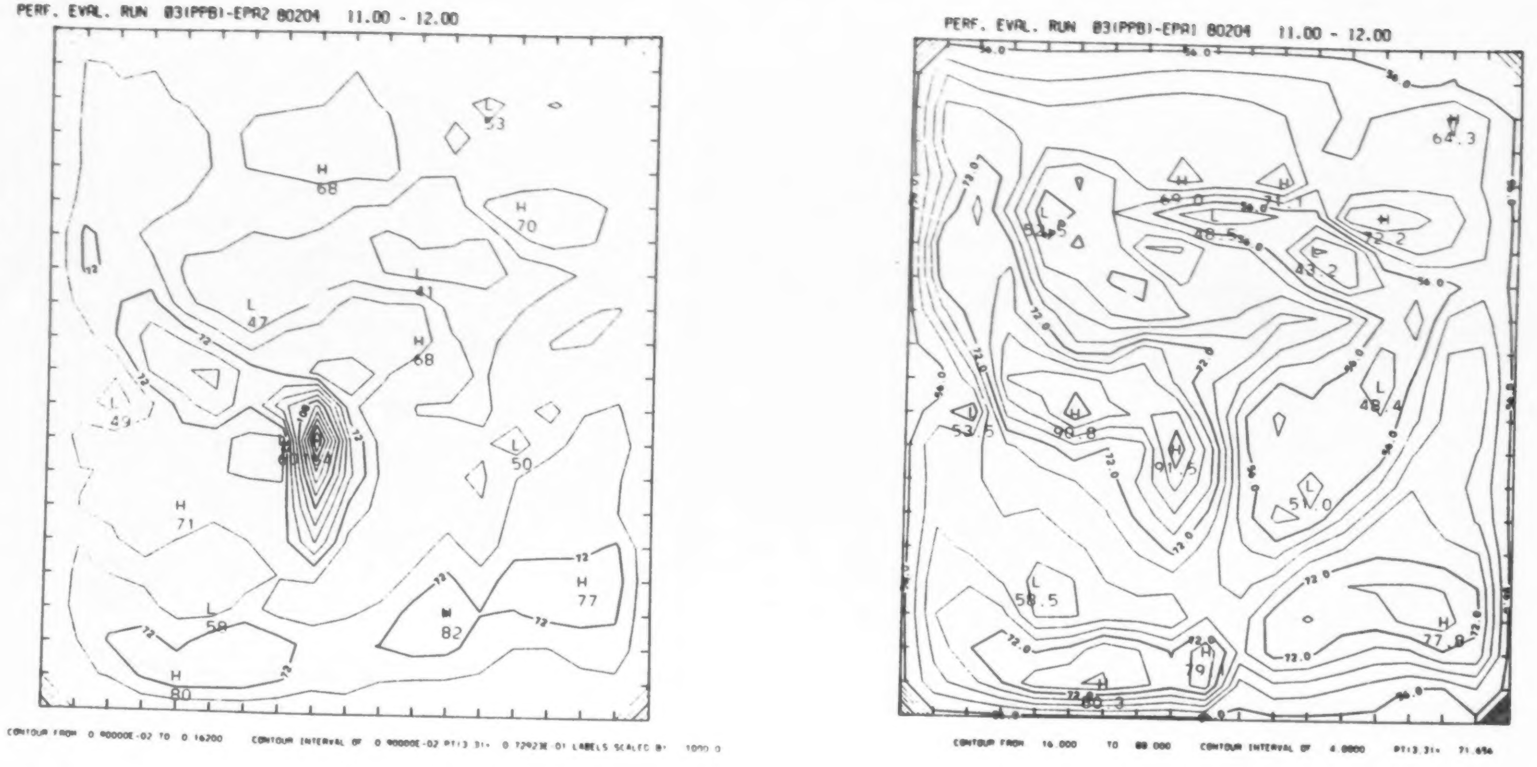


FIGURE 23: Comparison of the Simulations of EPA2 (left) with EPA1 (right)
Day 80204, Hour 12

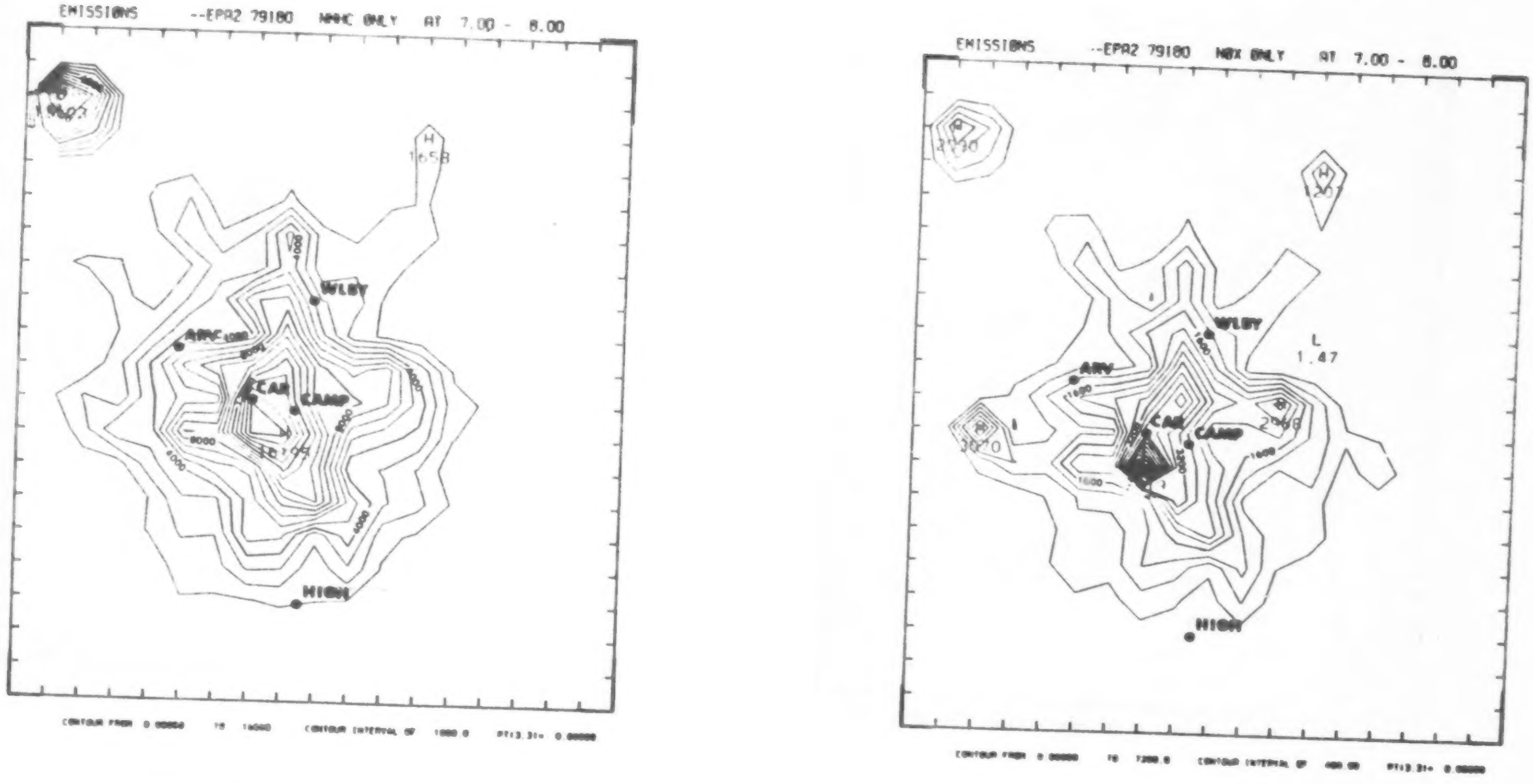

EMISSIONS --EPA2 79180 NMHC ONLY AT 7.00 - 8.00
H
1658
WLBY
ARV
CAR
CAMP
HIGH
4000
8000
CONTOUR FROM 0.00000 TO 16000 CONTOUR INTERVAL OF 1000.0 PT(3,3)= 0.00000
EMISSIONS --EPA2 79180 NOX ONLY AT 7.00 - 8.00
2130
H
1207
L
1.47
2968
3070
WLBY
ARV
CAR
CAMP
HIGH
1600
3200
CONTOUR FROM 0.00000 TO 7200.0 CONTOUR INTERVAL OF 400.00 PT(3,3)= 0.00000

FIGURE 24: Contour Plots of Reactive Hydrocarbon (left) and NO_x (right) Ground-level Emissions

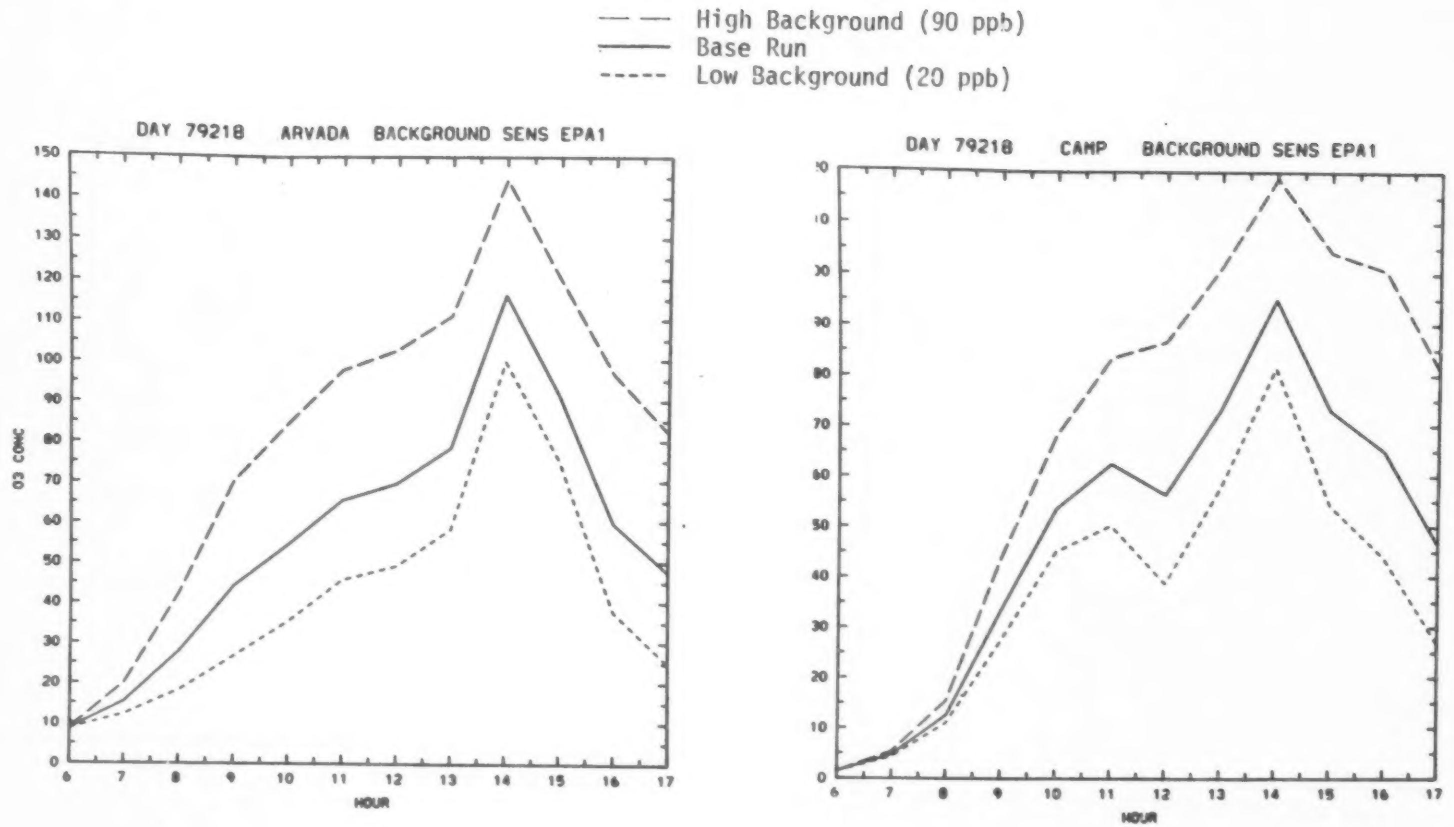


FIGURE 25 : Hourly Results of the Ozone Background Sensitivity Test for Each Monitoring Station: Day 79218.

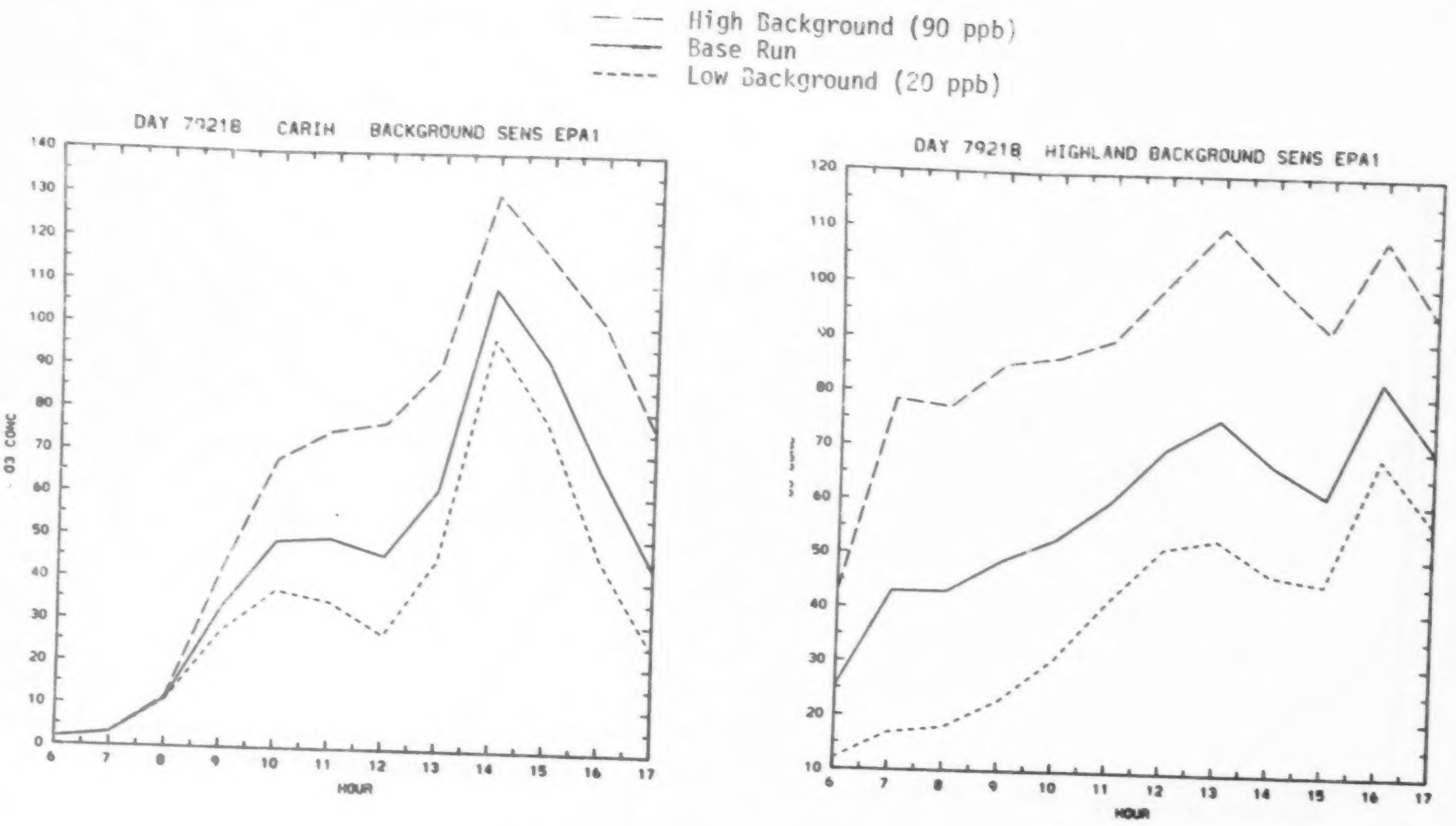


FIGURE 25, (Continued)

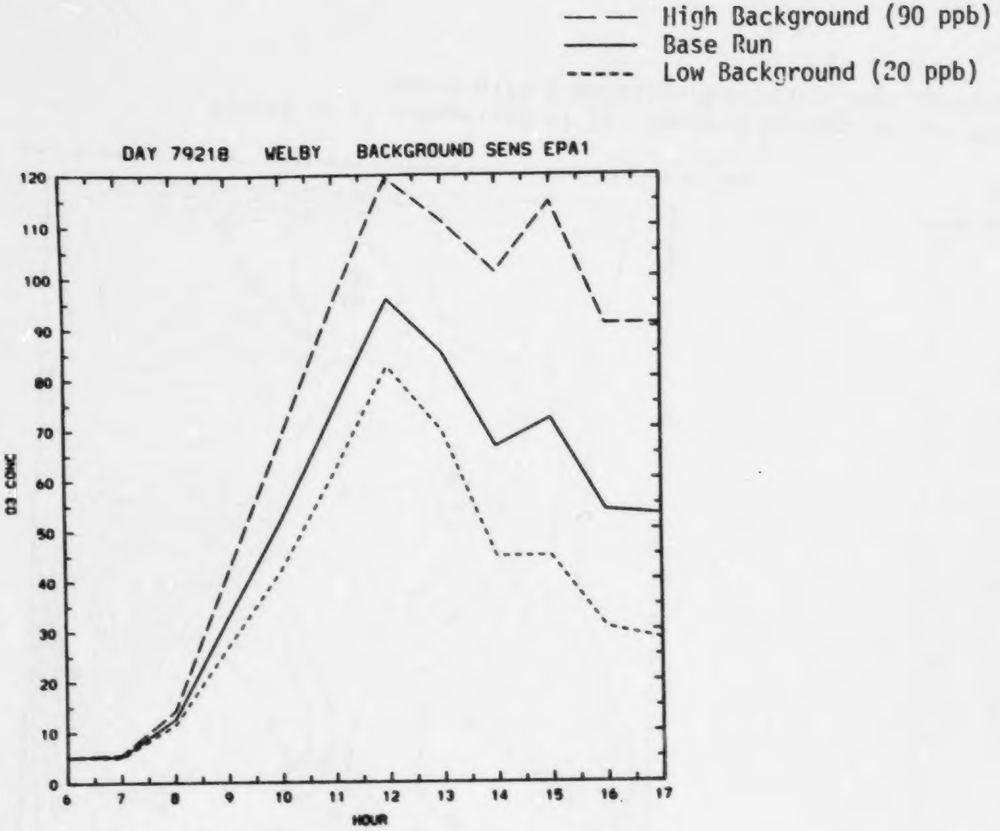


FIGURE 25, (Continued)

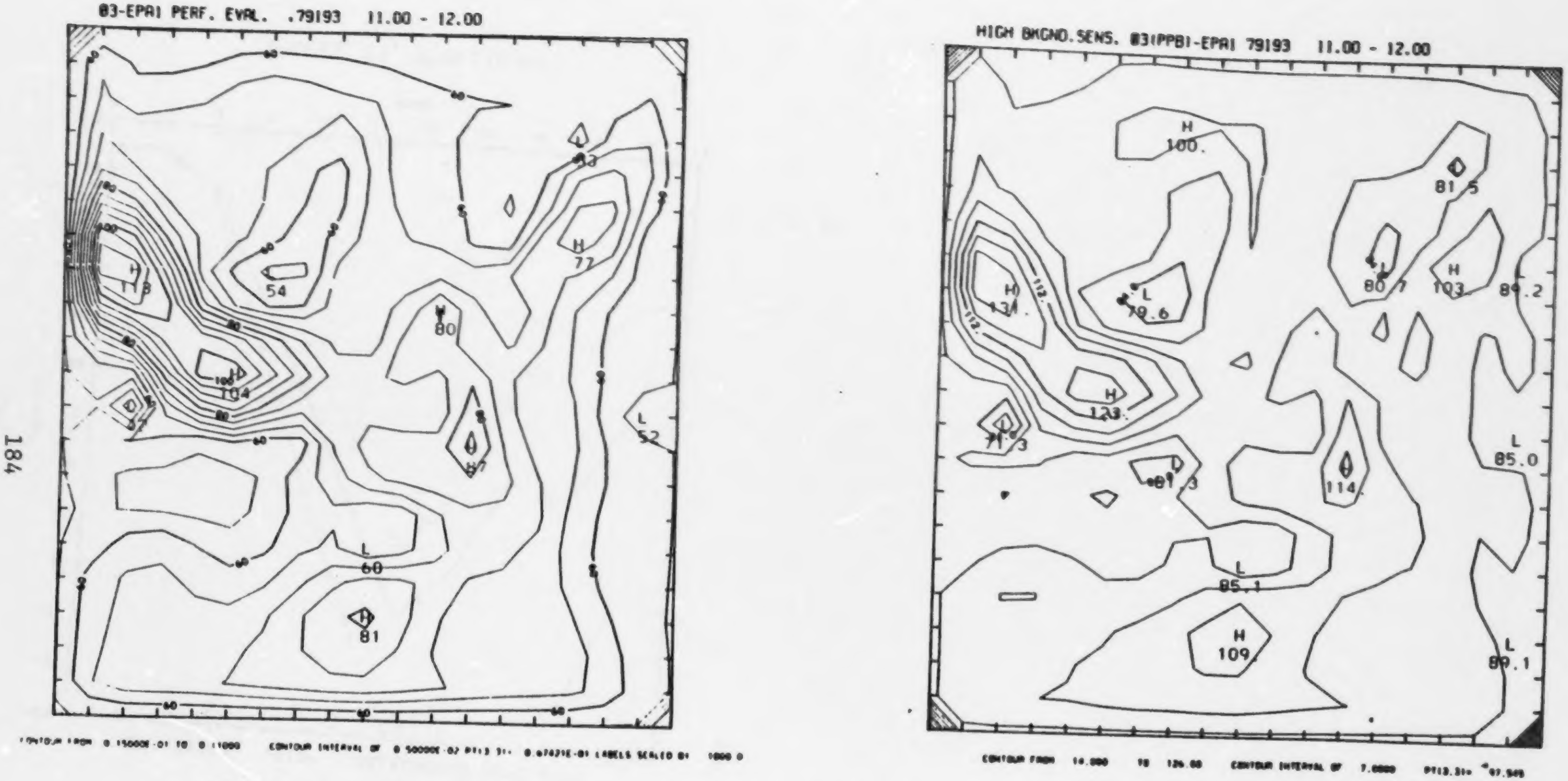

03-EPRI PERF. EVAL. .79193 11.00 - 12.00
60
H
113
54
104
37
80
H
77
33
87
L
52
L
60
H
81
HIGH BKGND.SENS. 03(PPB)-EPRI 79193 11.00 - 12.00
H
100.
H
131.
112.
L
79.6
H
123.
71.3
81.3
81.5
80.7
H
103.
89.2
114.
L
85.0
L
85.1
H
109.
L
89.1
CONTOUR FROM 14.000 TO 126.00 CONTOUR INTERVAL OF 7.0000

FIGURE 26 : Comparison of Performance Evaluation Run Using Regular Background Ozone with Simulation Run Using High Background Ozone

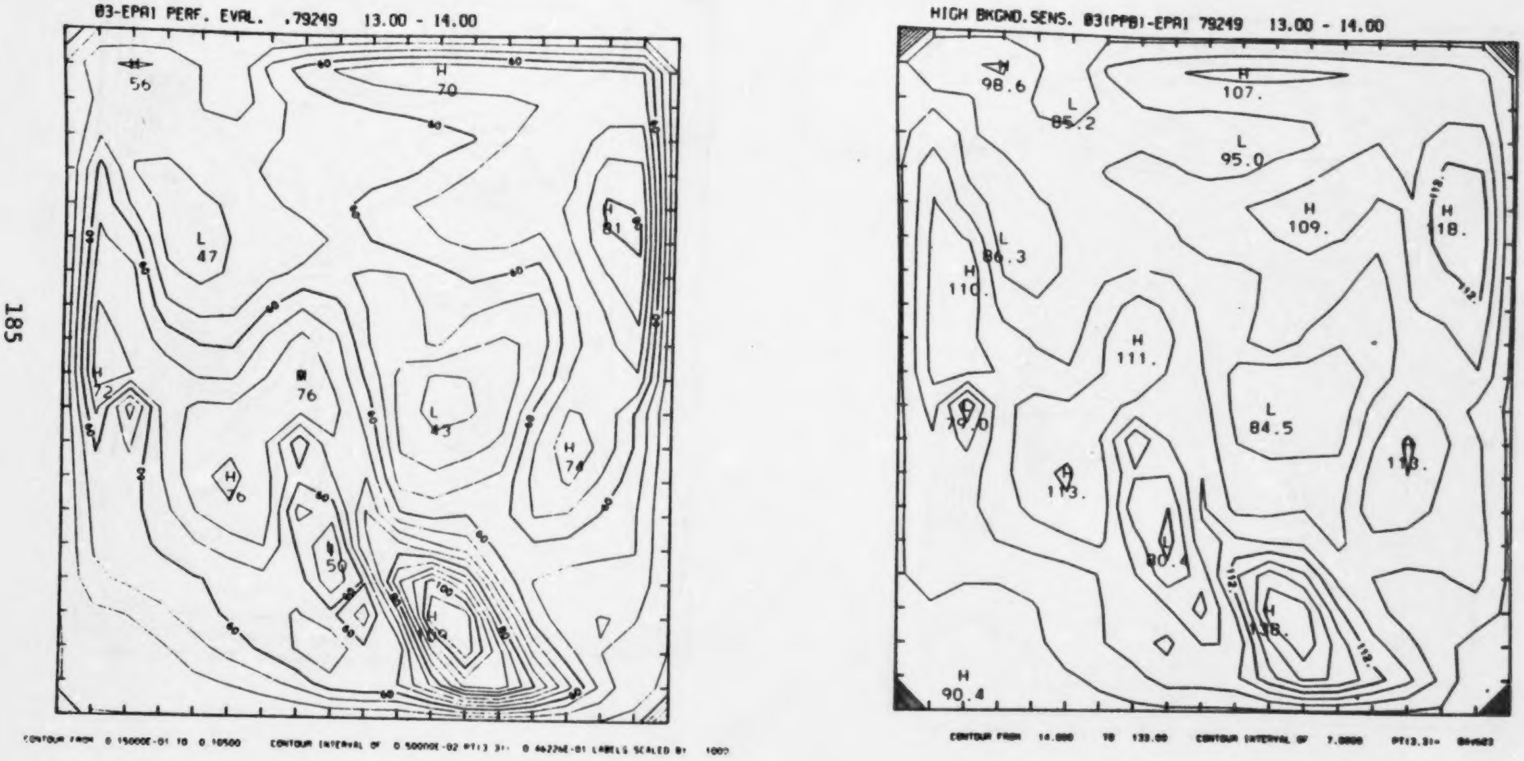


FIGURE 27 : Comparison of Performance Evaluation Run Using Regular Background Ozone with Simulation Run Using High Background Ozone

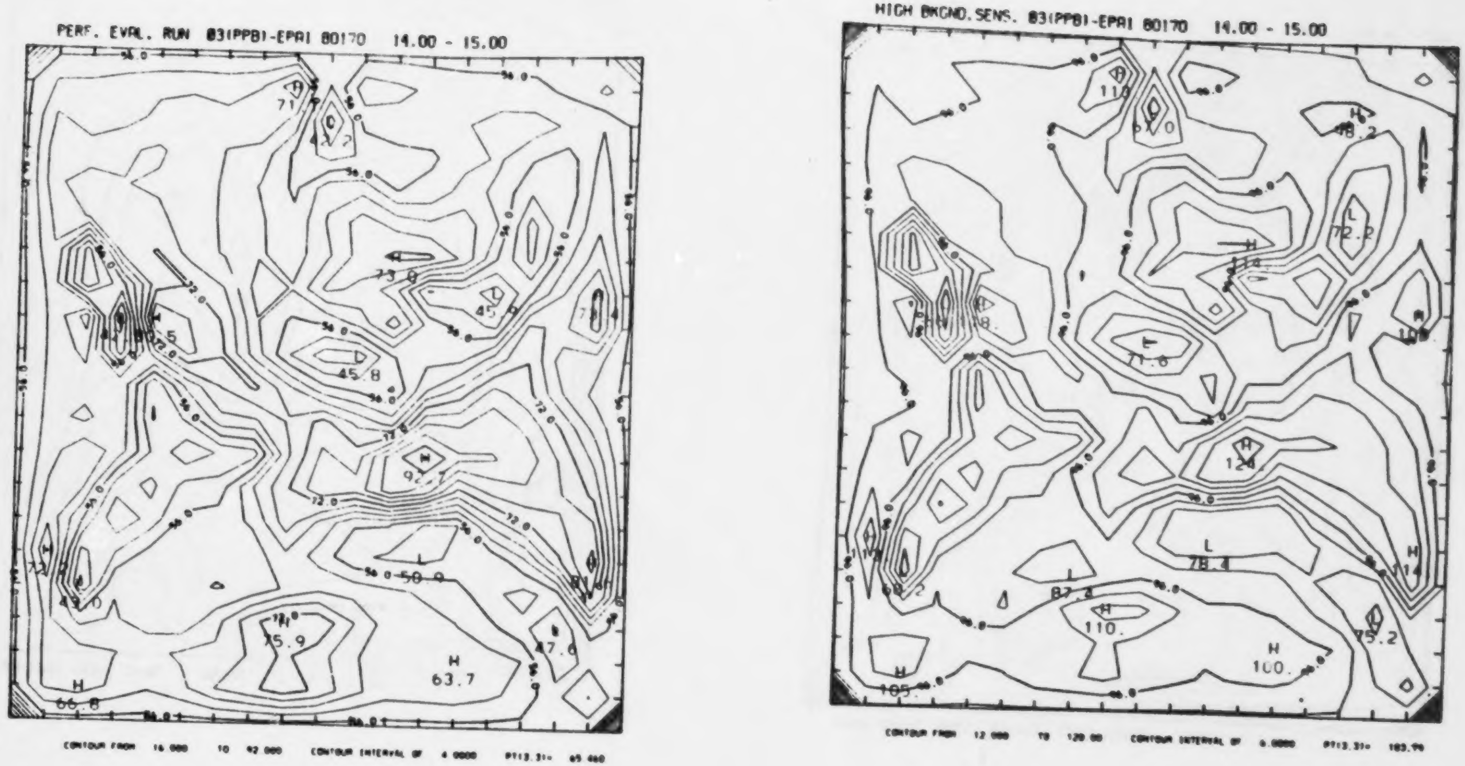

PERF. EVAL. RUN 03(PPB)-EPRI 80170 14.00 - 15.00
H
71.
42.2
73.0
L
45.9
45.8
H
92.7
72.0
56.0
50.9
75.9
H
63.7
47.6
H
66.8
43.0
CONTOUR FROM 16.000 TO 92.000 CONTOUR INTERVAL OF 4.0000 PT(3,3)= 65.460
HIGH BKGND.SENS. 03(PPB)-EPRI 80170 14.00 - 15.00
H
110
97.0
H
98.2
L
72.2
H
114
L
71.6
H
109
H
124
L
78.4
H
114
87.4
H
110.
H
100.
75.2
H
105.
CONTOUR FROM 12.000 TO 120.00 CONTOUR INTERVAL OF 6.0000 PT(3,3)= 103.79

FIGURE 28 : Comparison of Performance Evaluation Run Using Regular Background Ozone with Simulation Run Using High Background Ozone

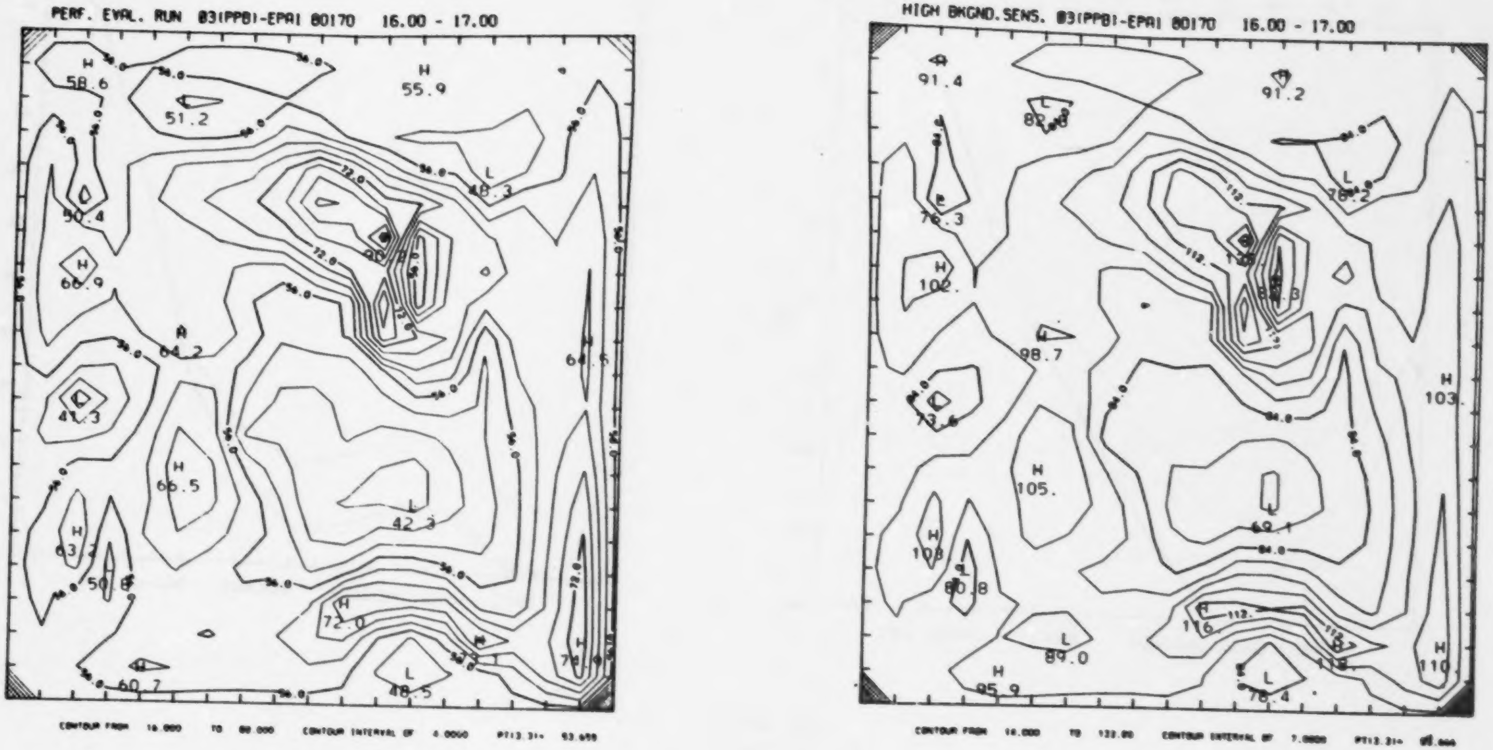


FIGURE 29 : Comparison of Performance Evaluation Run Using Regular Background Ozone with Simulation Run Using High Background Ozone

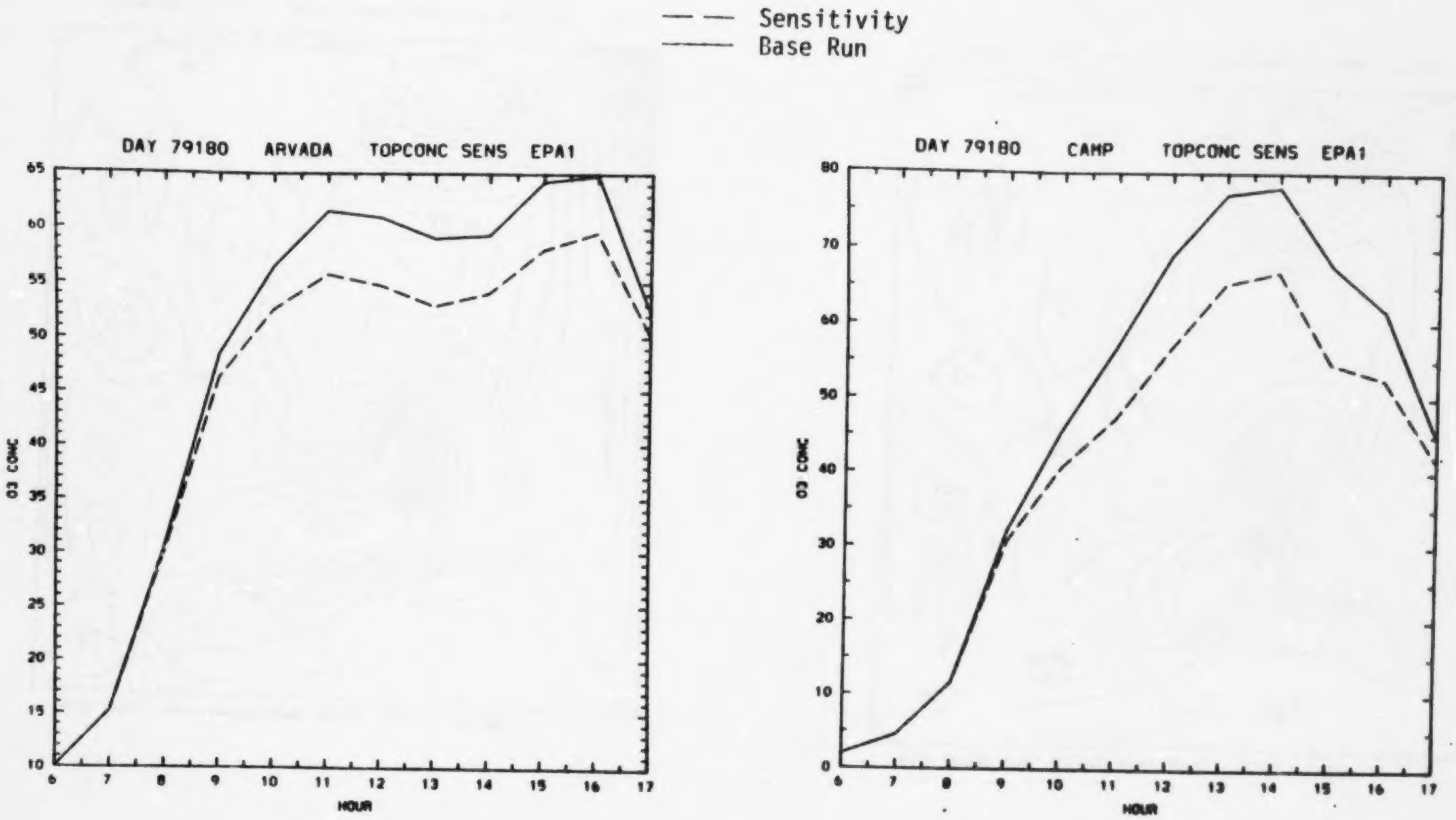


FIGURE 30 : Hourly Results of Hydrocarbon Background Sensitivity Test for Each Monitoring Station.

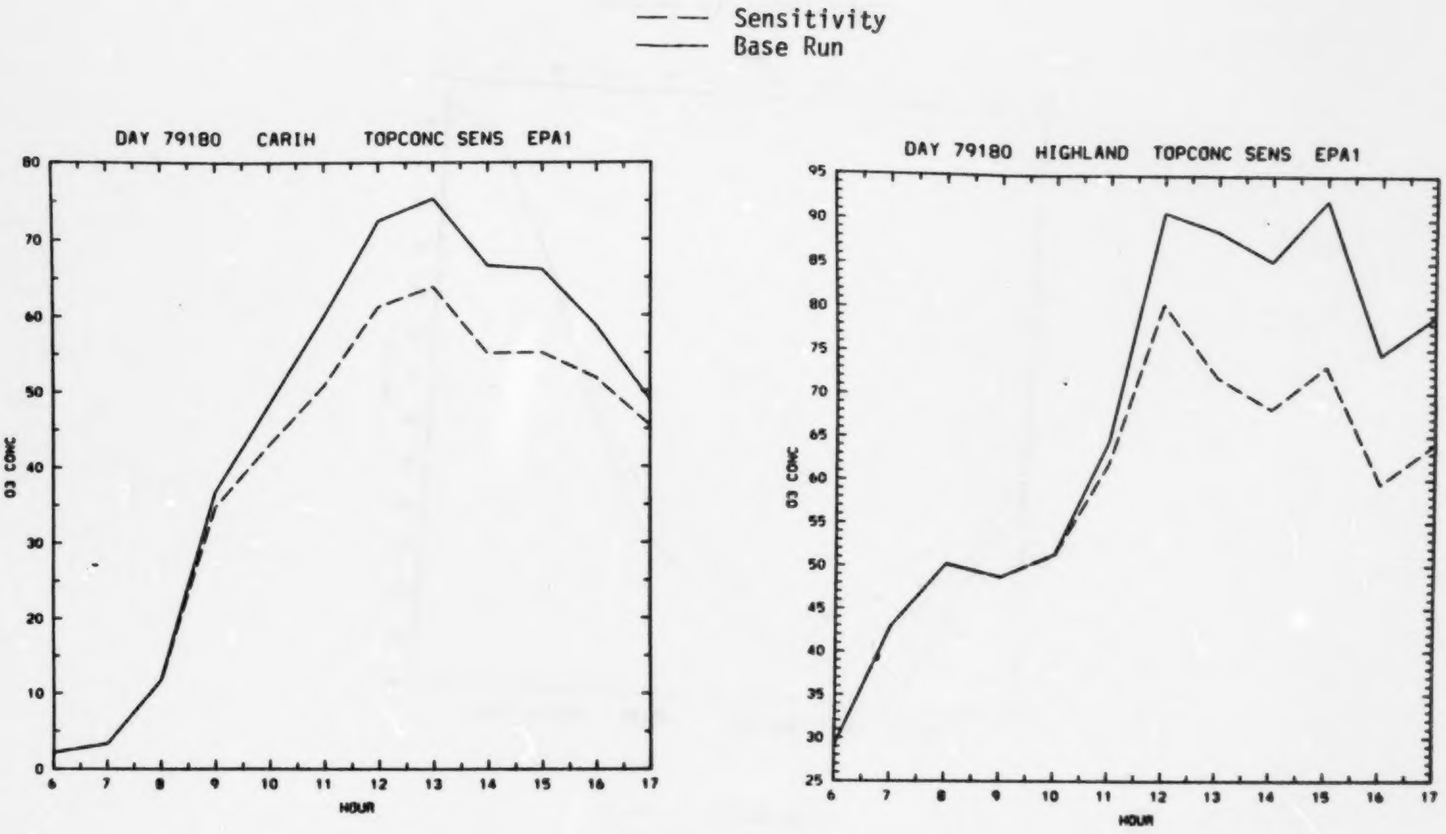


FIGURE 30, (Continued)

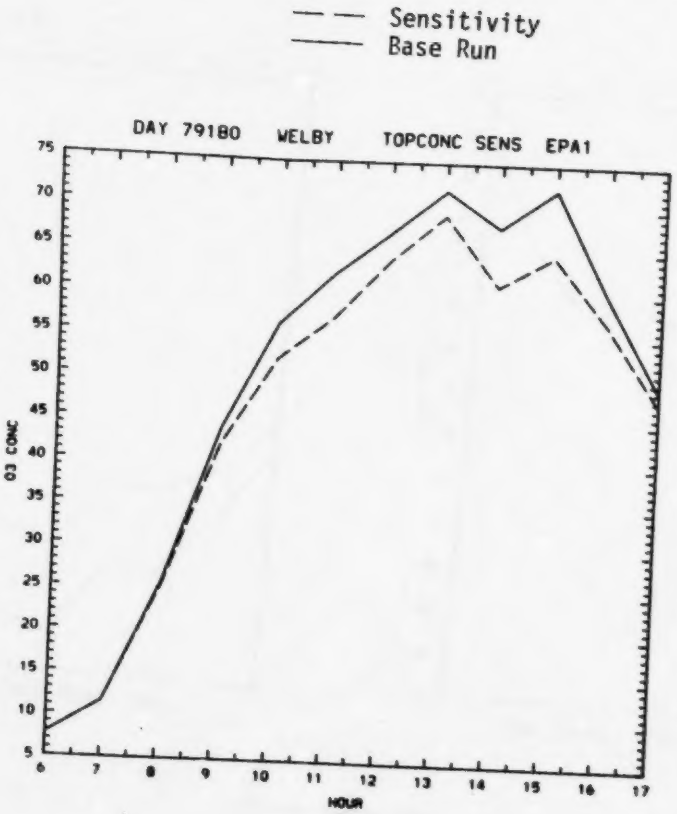


FIGURE 30, (continued)

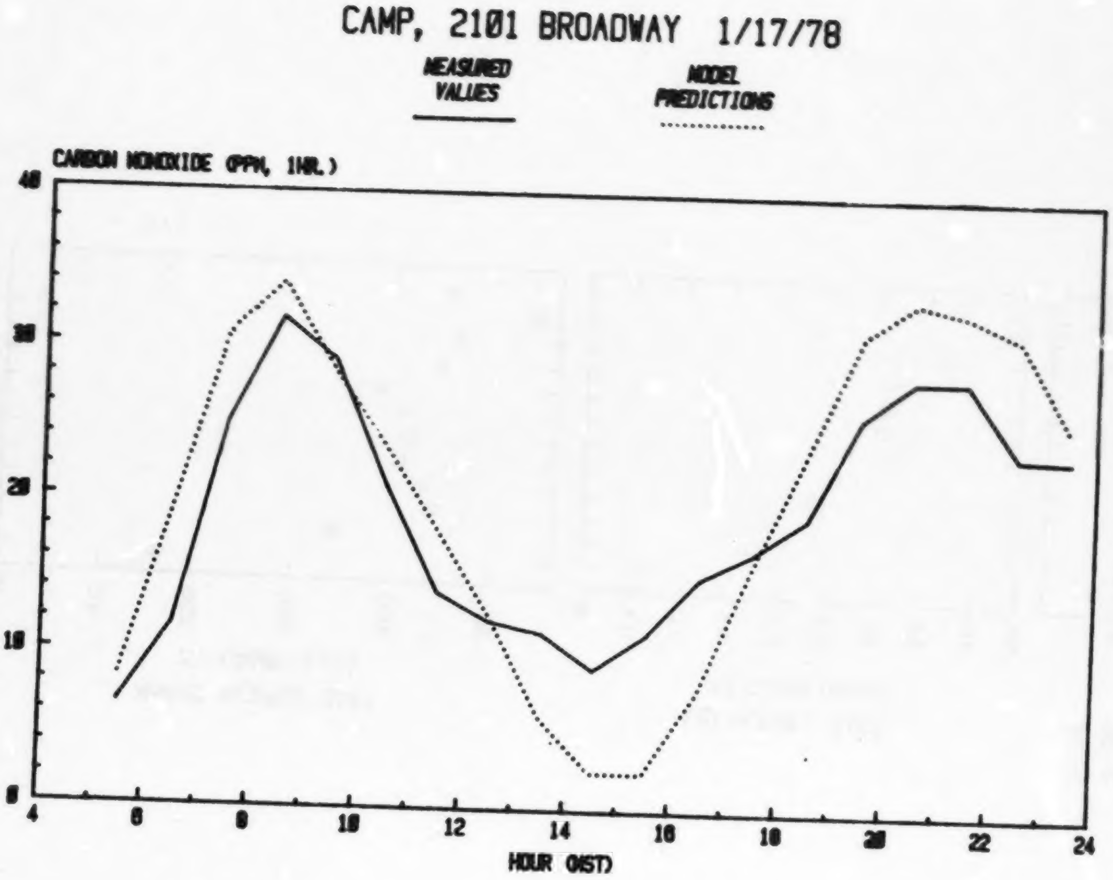


FIGURE 31 : Hourly Predicted and Observed Carbon Monoxide Concentrations at CAMP for an Extreme Winter Day.

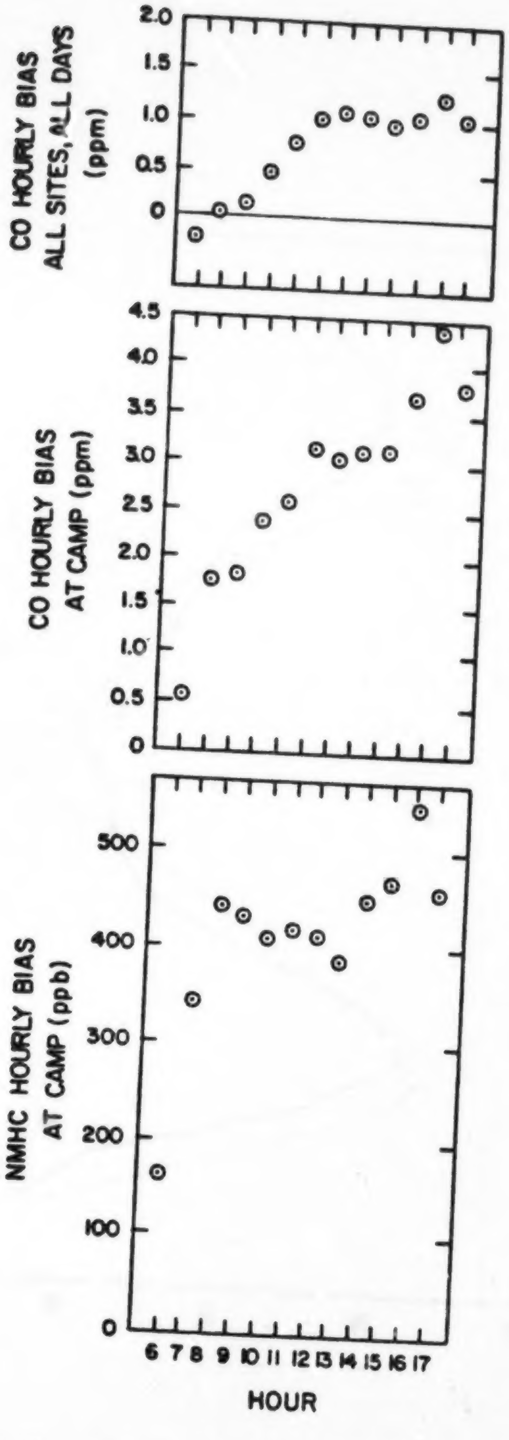


FIGURE 32 : Hourly Bias for Carbon Monoxide (CO) and Nonmethane Hydrocarbons (NMHC) Averaged over the 11 Days.

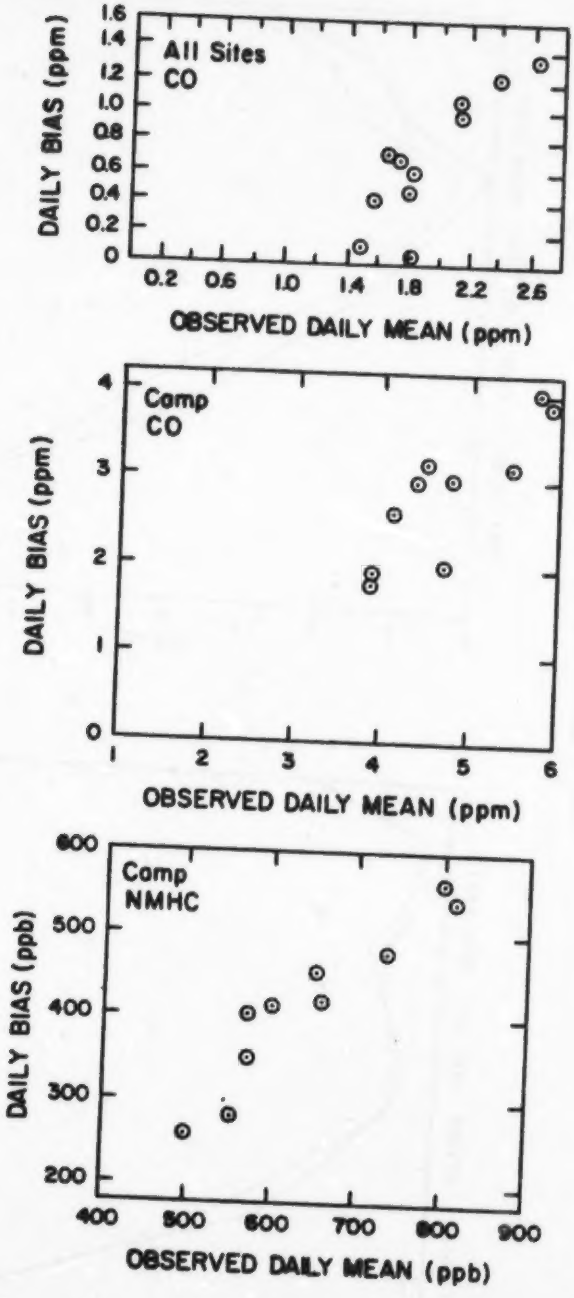


FIGURE 33 : Daily Mean Observed Concentrations Compared with Daily Bias for Carbon Monoxide (CO) and Nonmethane Hydrocarbons (NMHC).

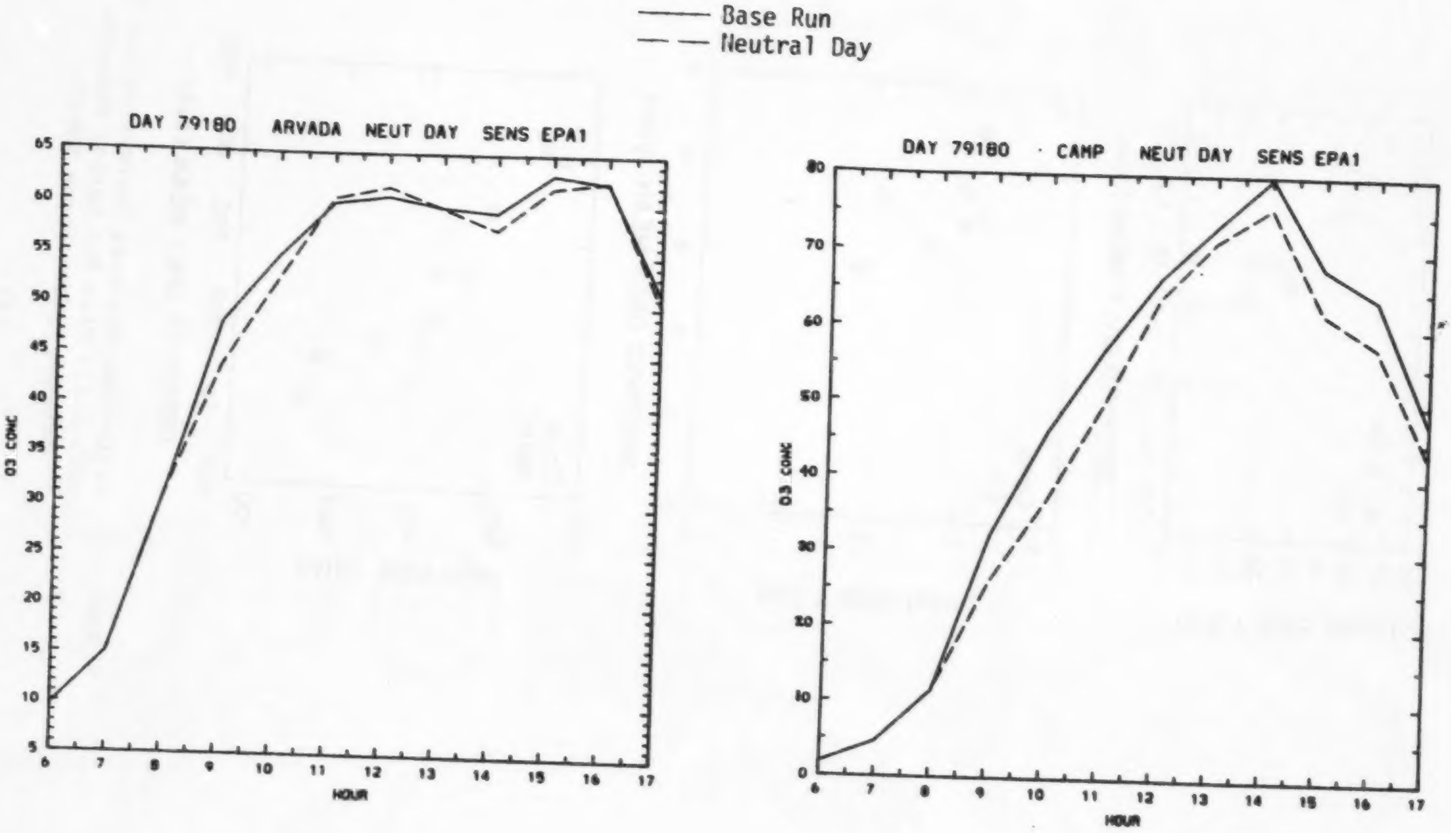


FIGURE 34: Hourly Results of the Neutral Day Sensitivity with EPA1 for Ozone on 79180 at Each Monitoring Site.

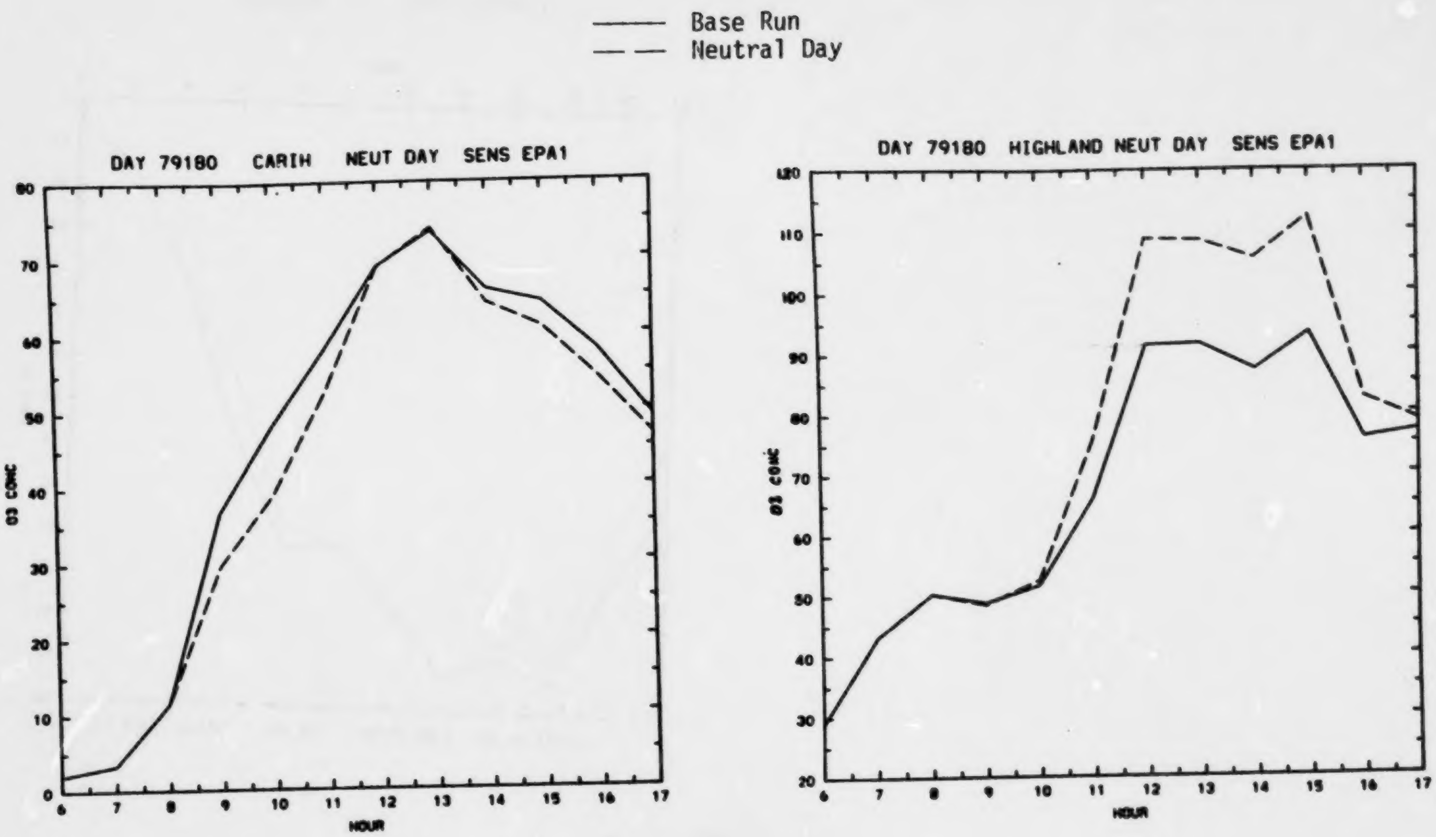


FIGURE 34. (Continued)

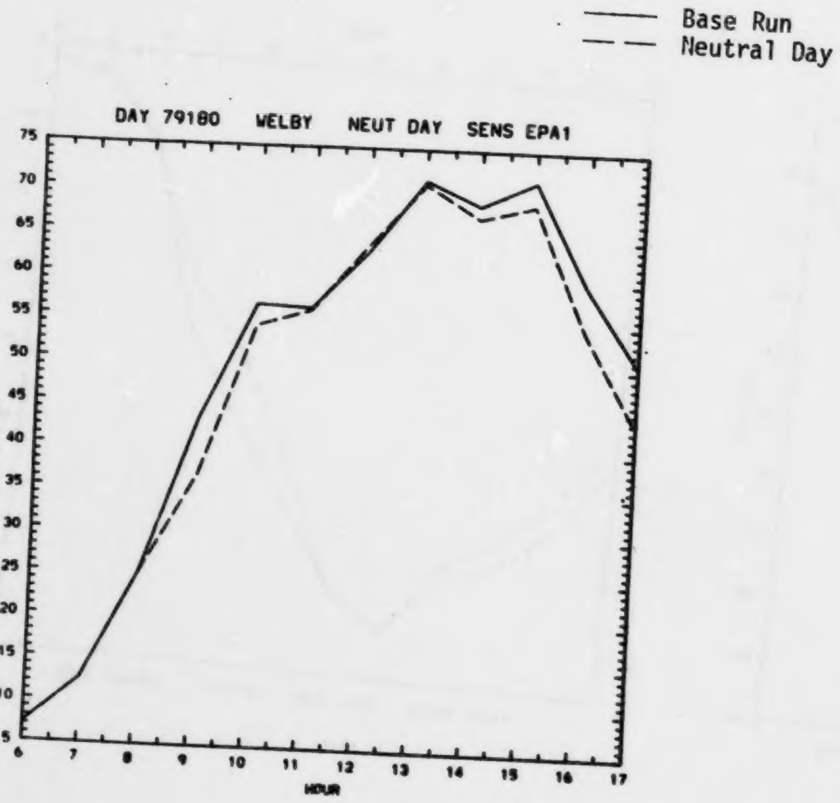


FIGURE 34. (Continued)

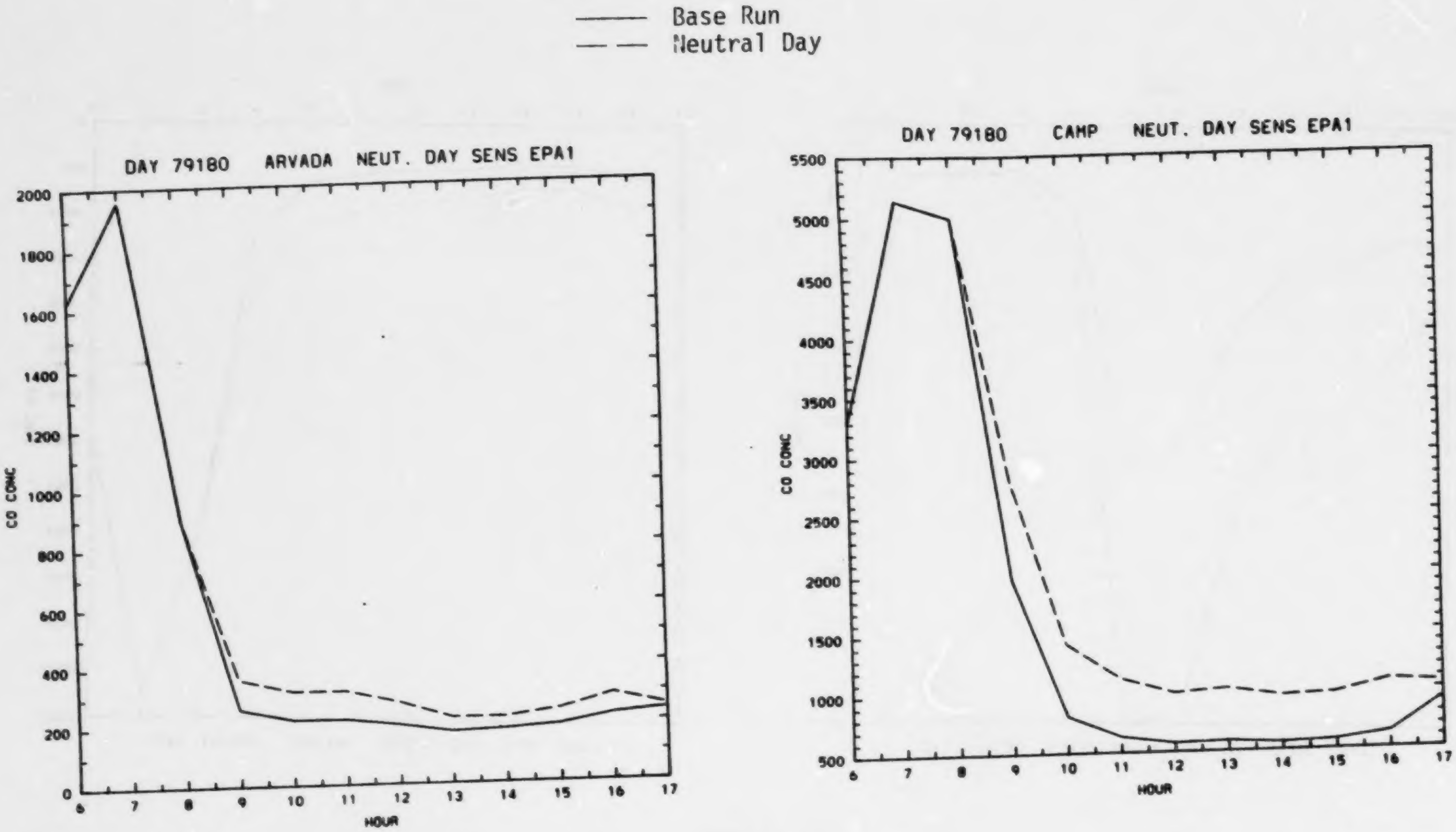


FIGURE 35: Hourly Results of the Neutral Day Sensitivity in EPA1 for Carbon Monoxide on 79180 at Each Monitoring Site.

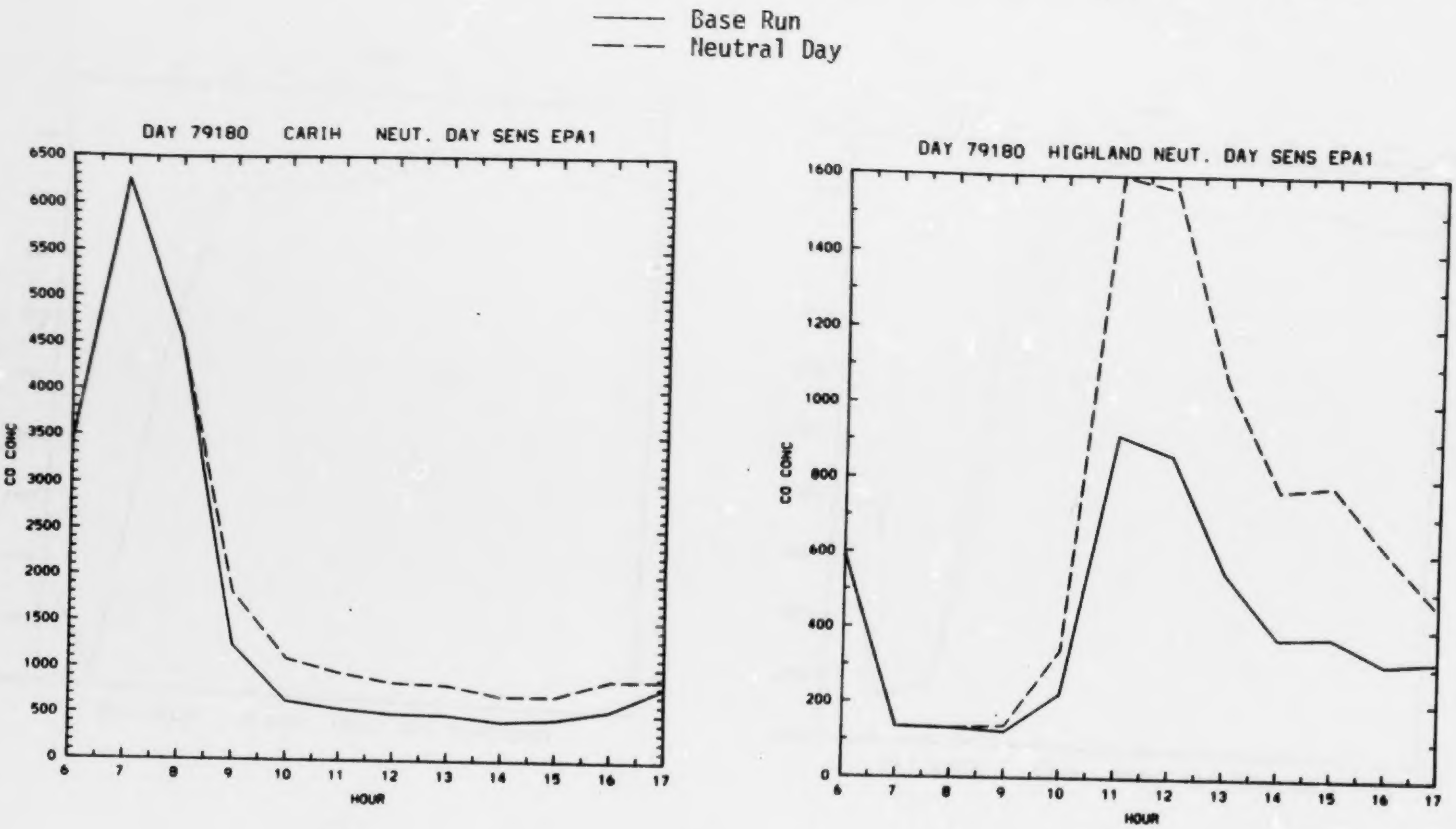


FIGURE 35, (Continued)

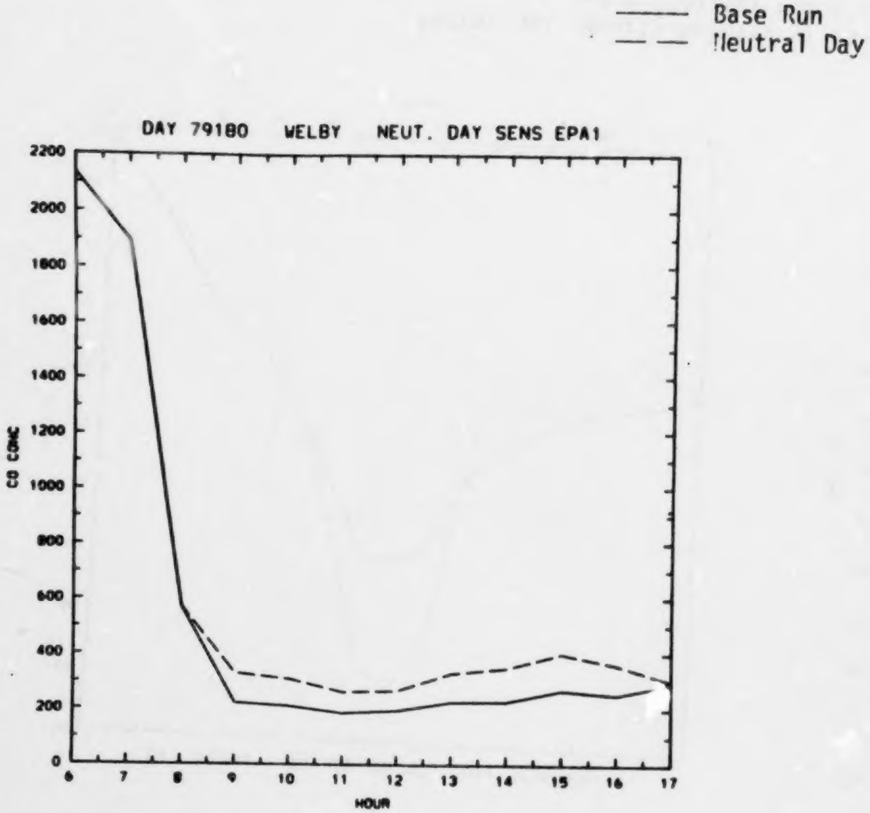


FIGURE 35, (Continued)

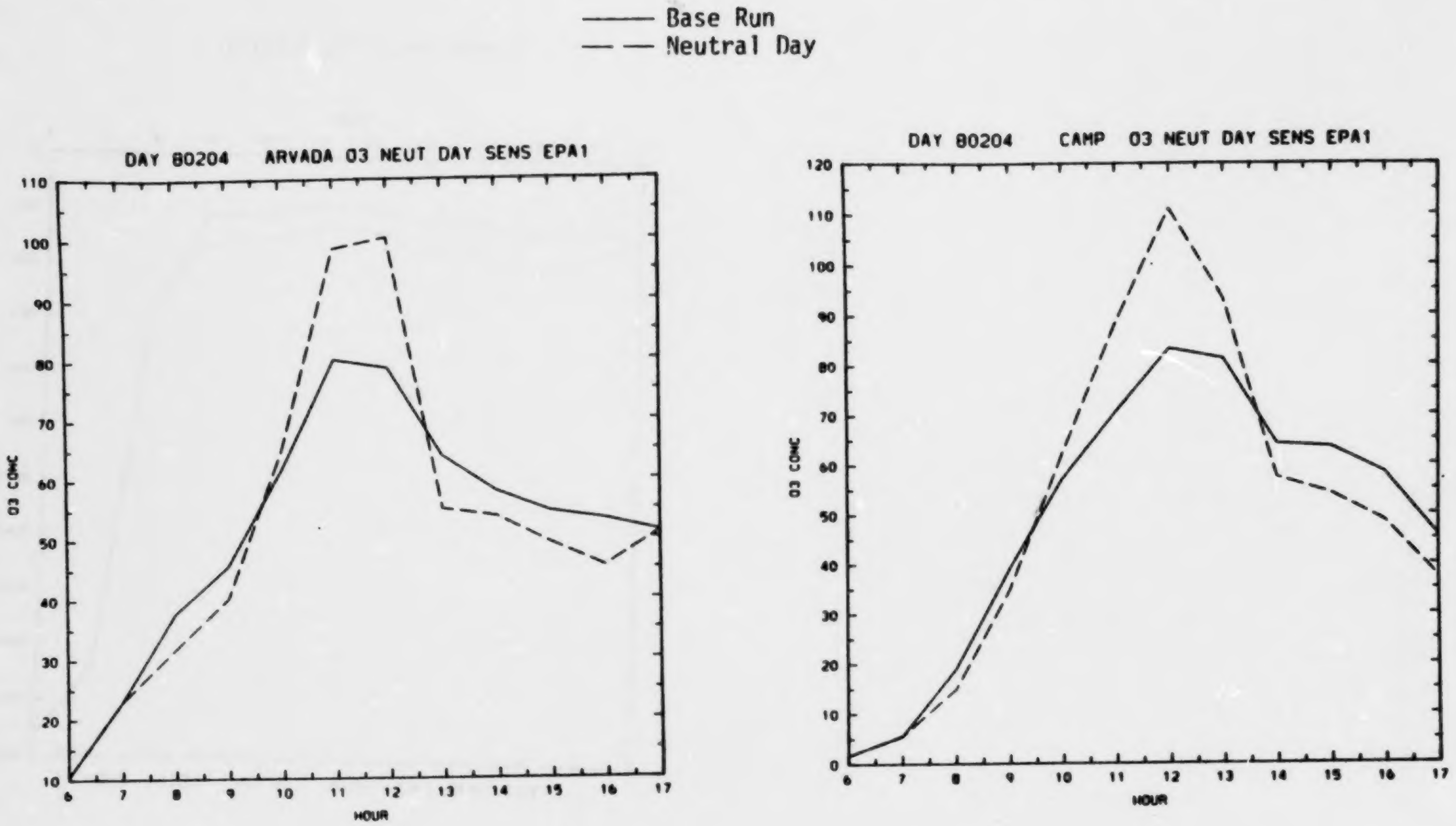


FIGURE 36: Hourly Results of the Neutral Day Sensitivity with EPA1 for Ozone on 80204 at Each Monitoring Site.

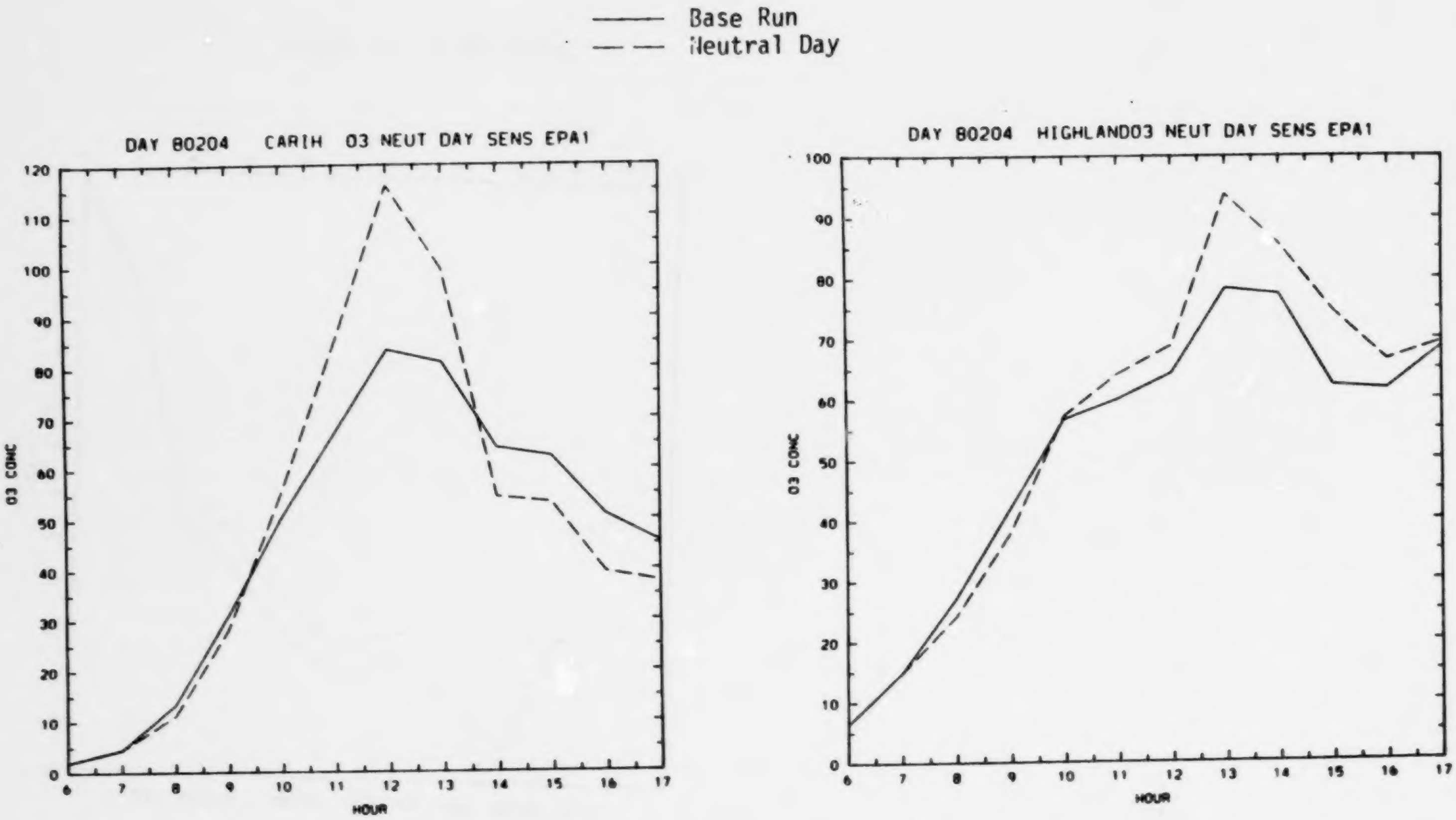


FIGURE 36, (Continued)

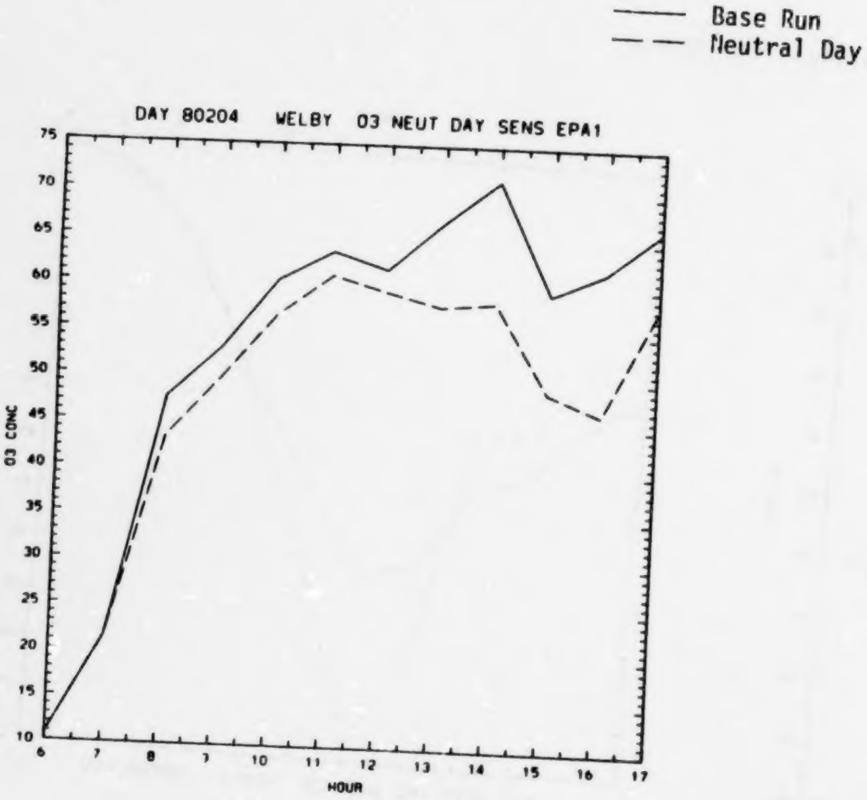


FIGURE 36, (Continued)

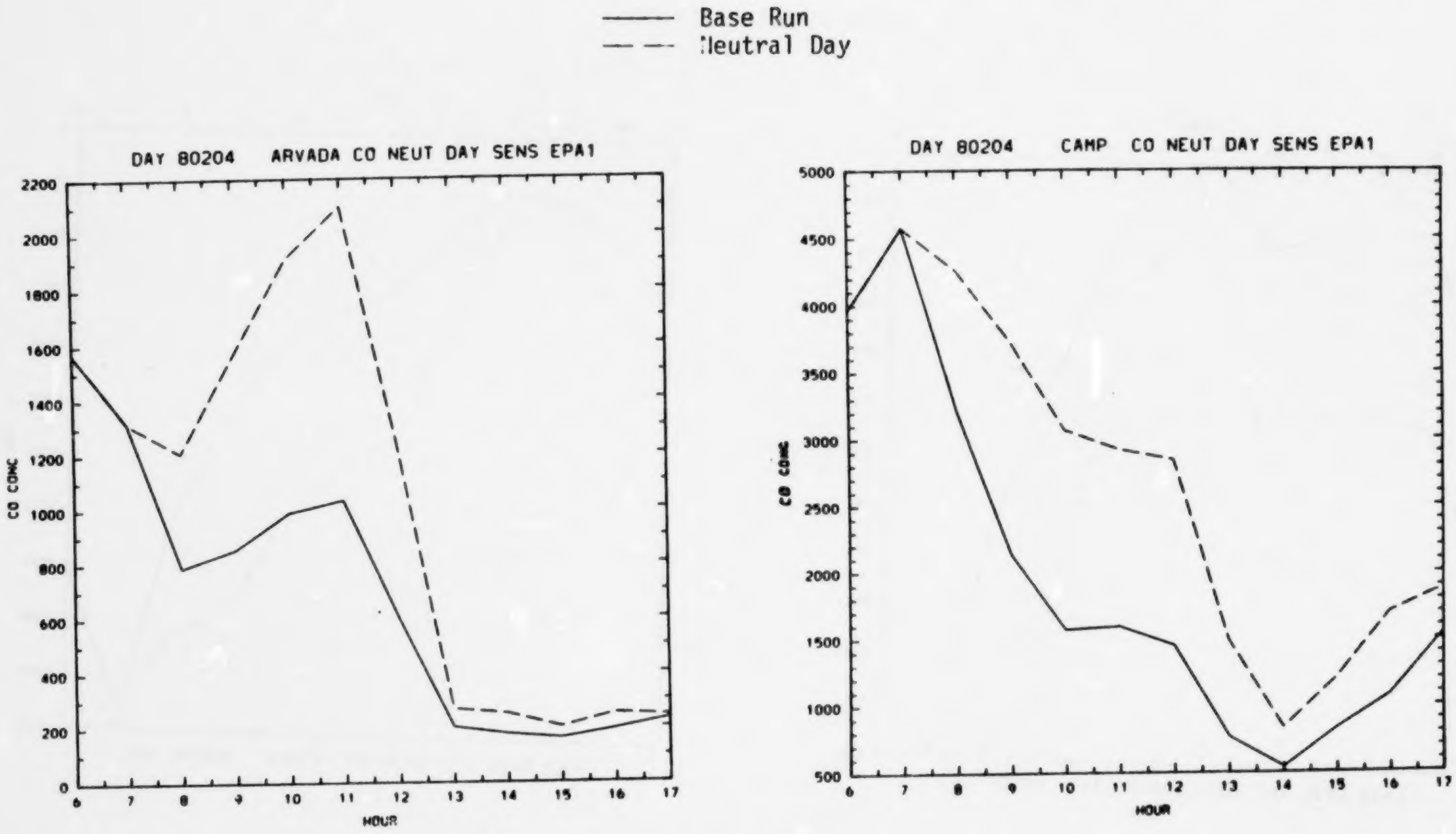


FIGURE 37: Hourly Results of the Neutral Day Sensitivity with EPA1 for Carbon Monoxide on 80204 at Each Monitoring Site.

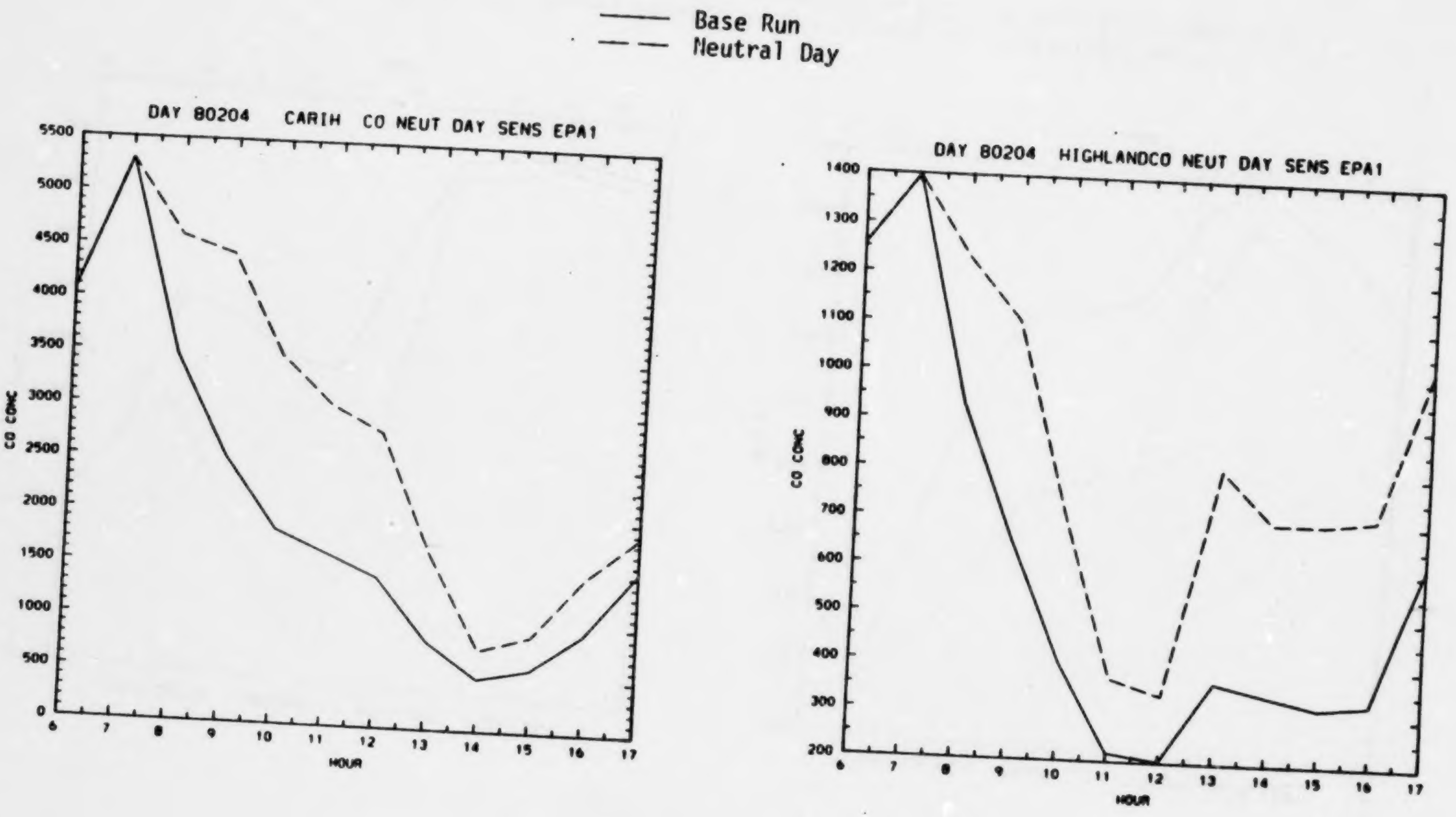


FIGURE 37. (Continued)

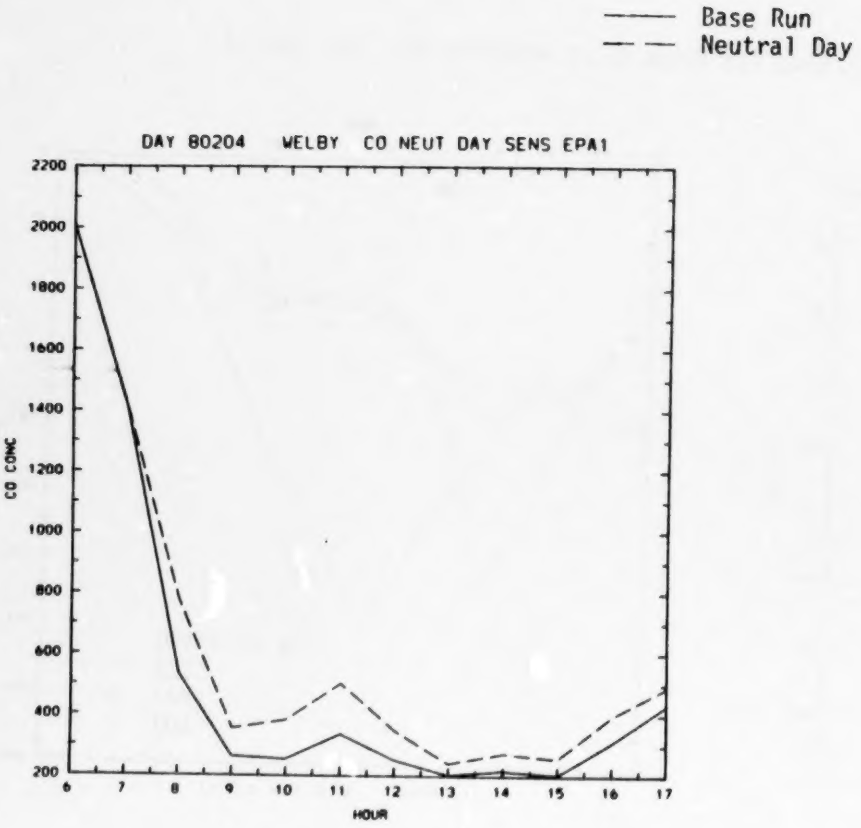


FIGURE 37, (Continued)

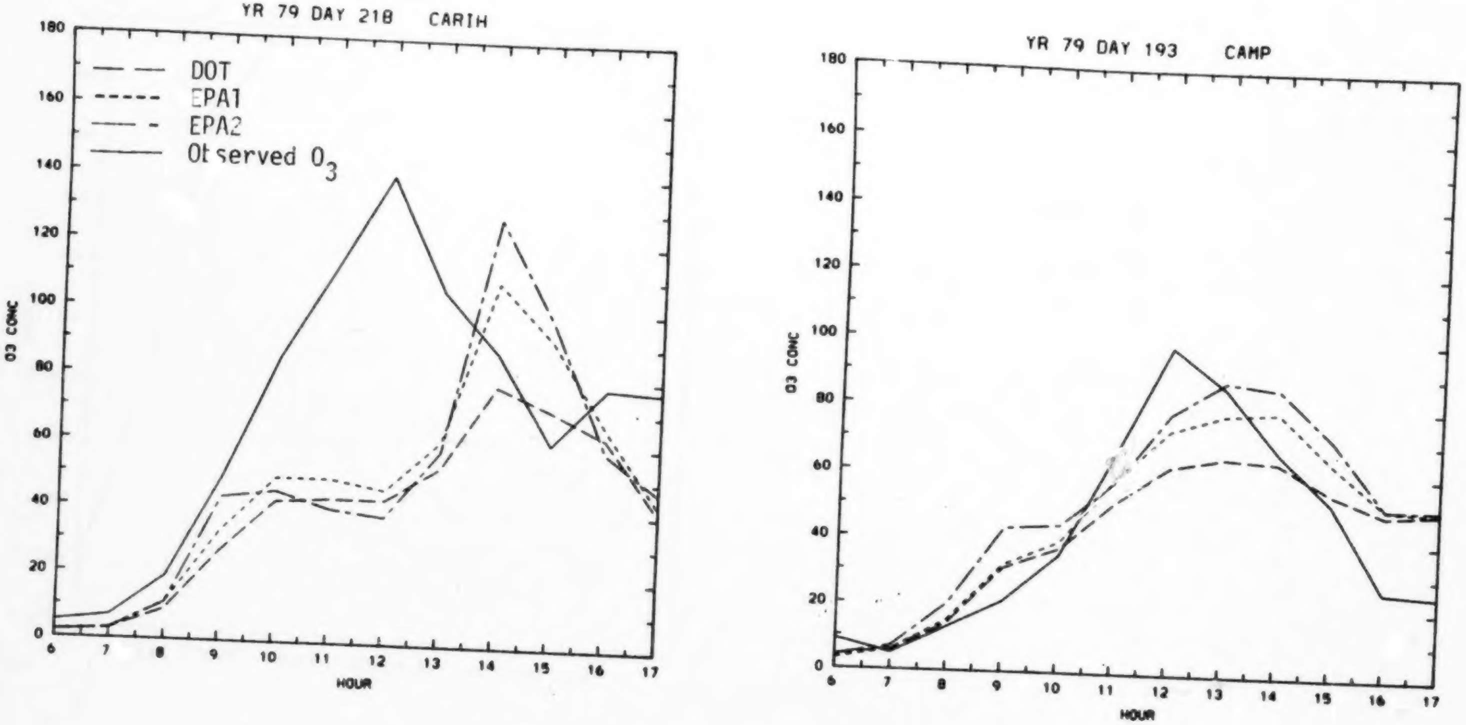
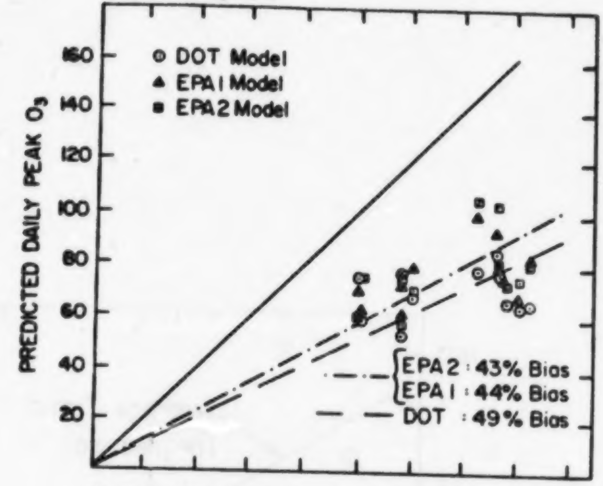
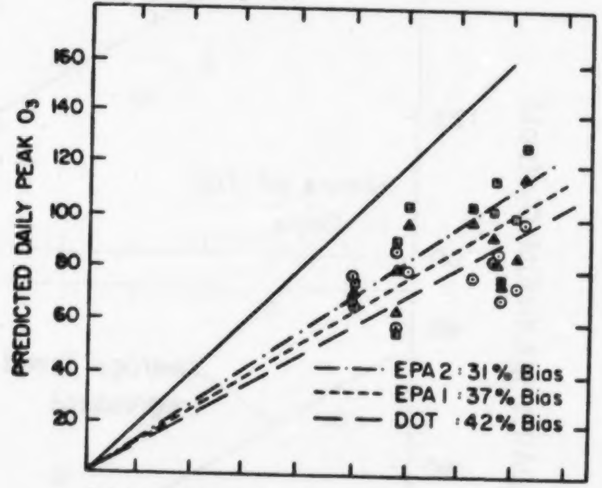


FIGURE 38: Two Examples of Missing the Peak in Time for the Three Models.

(a) Paired by Site



(b) Unpaired by Site



(c) Unconstrained in Space

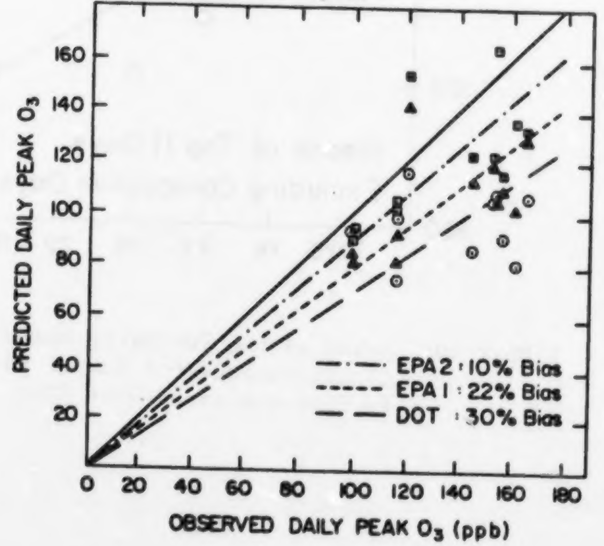


FIGURE 39: Predicted versus Observed Daily Maximum Concentrations under Three Pairing Methods.

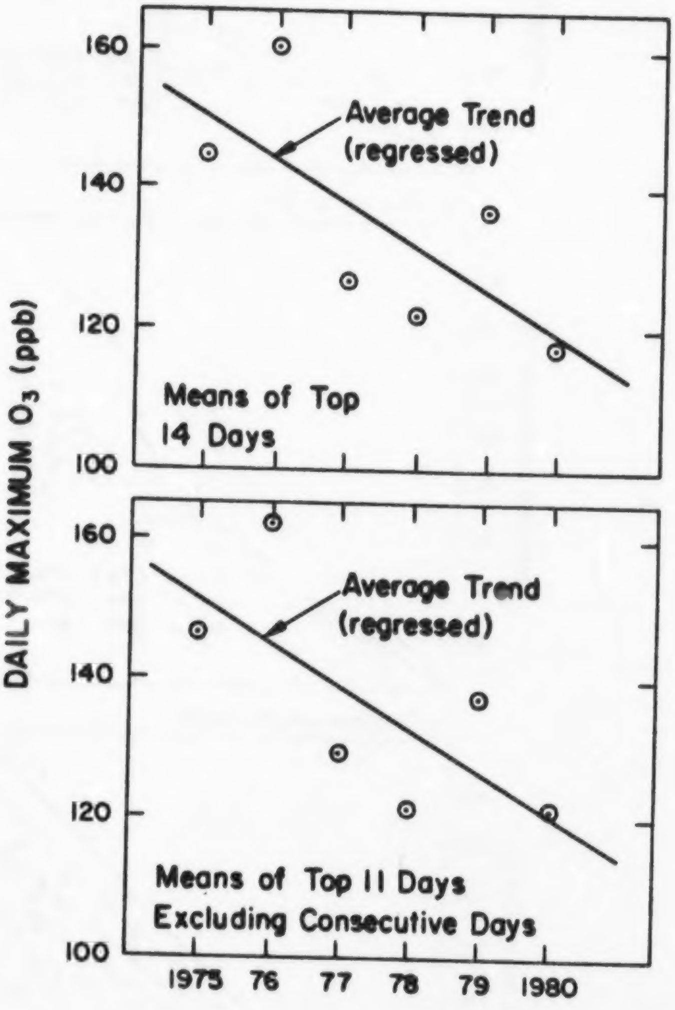


FIGURE 40: Means of the Top Daily Maximum Ozone Concentrations for Summer for 1975-1980 to Show the Trend Over Time.

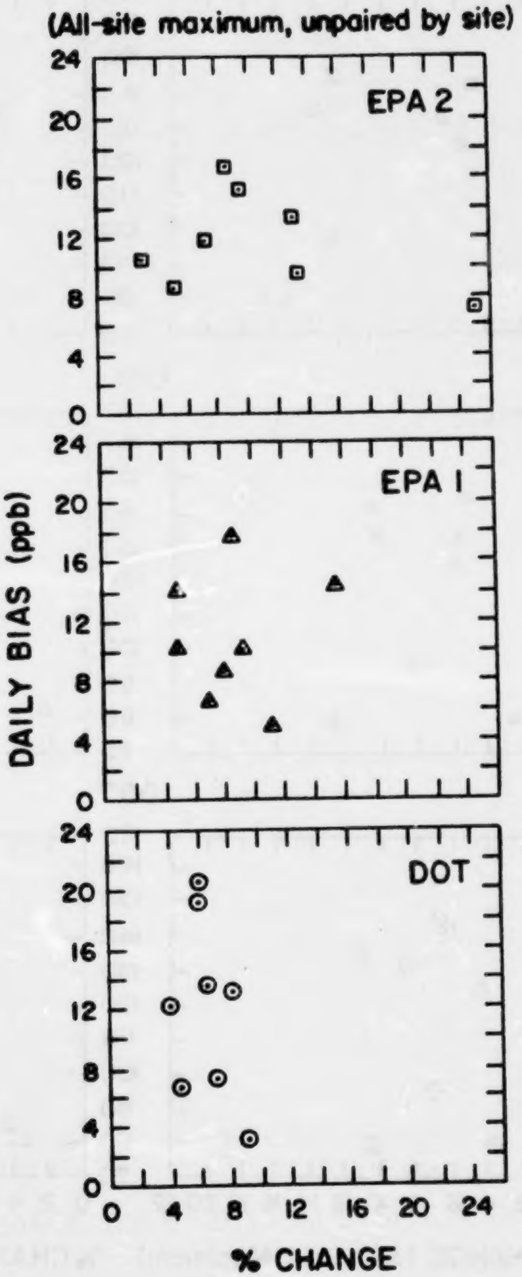


FIGURE 41: Emissions Change Results versus Daily Bias for the Three Models.

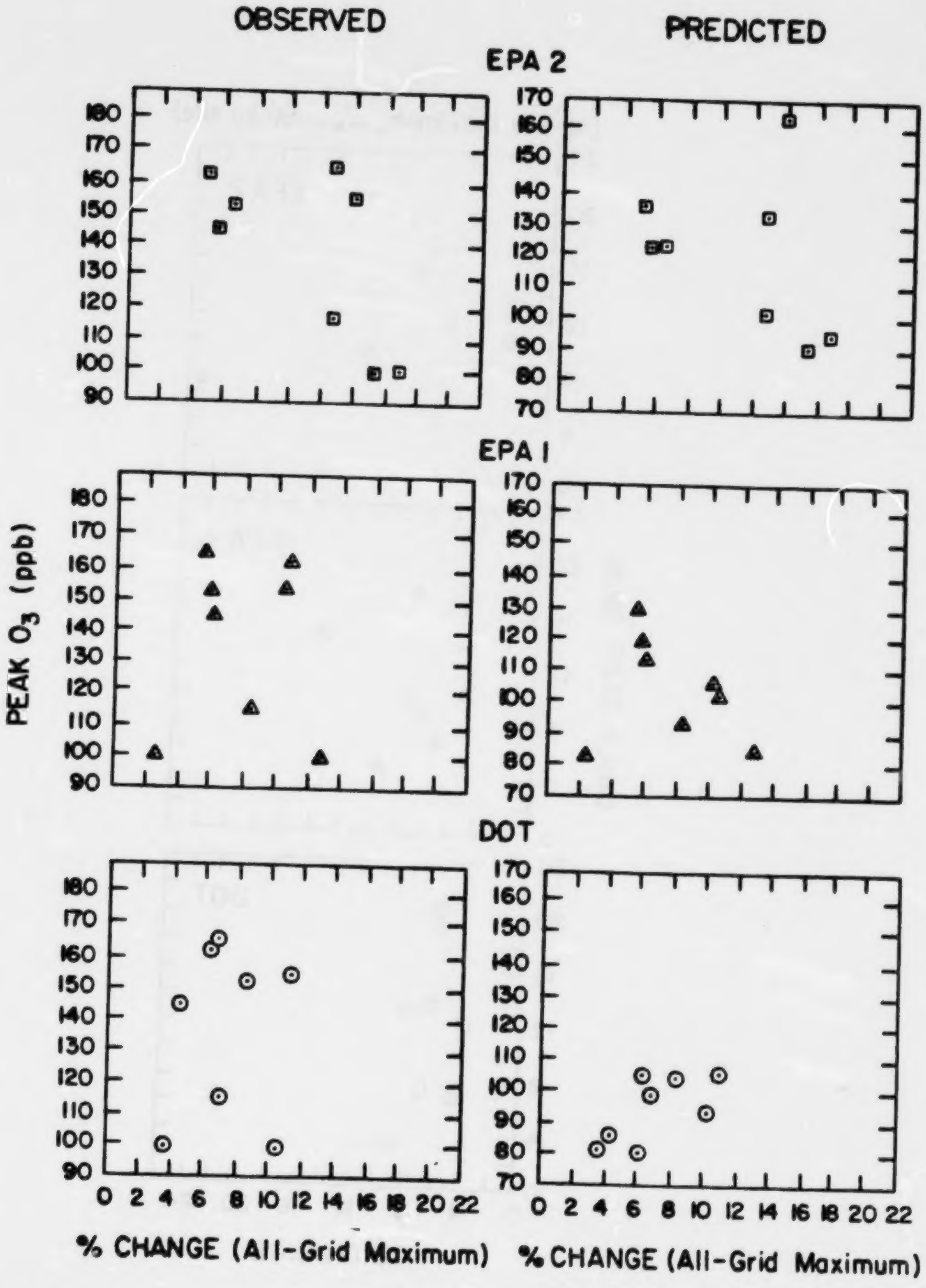


FIGURE 42: Emissions Change Results versus Peak Ozone--Observed and Predicted.

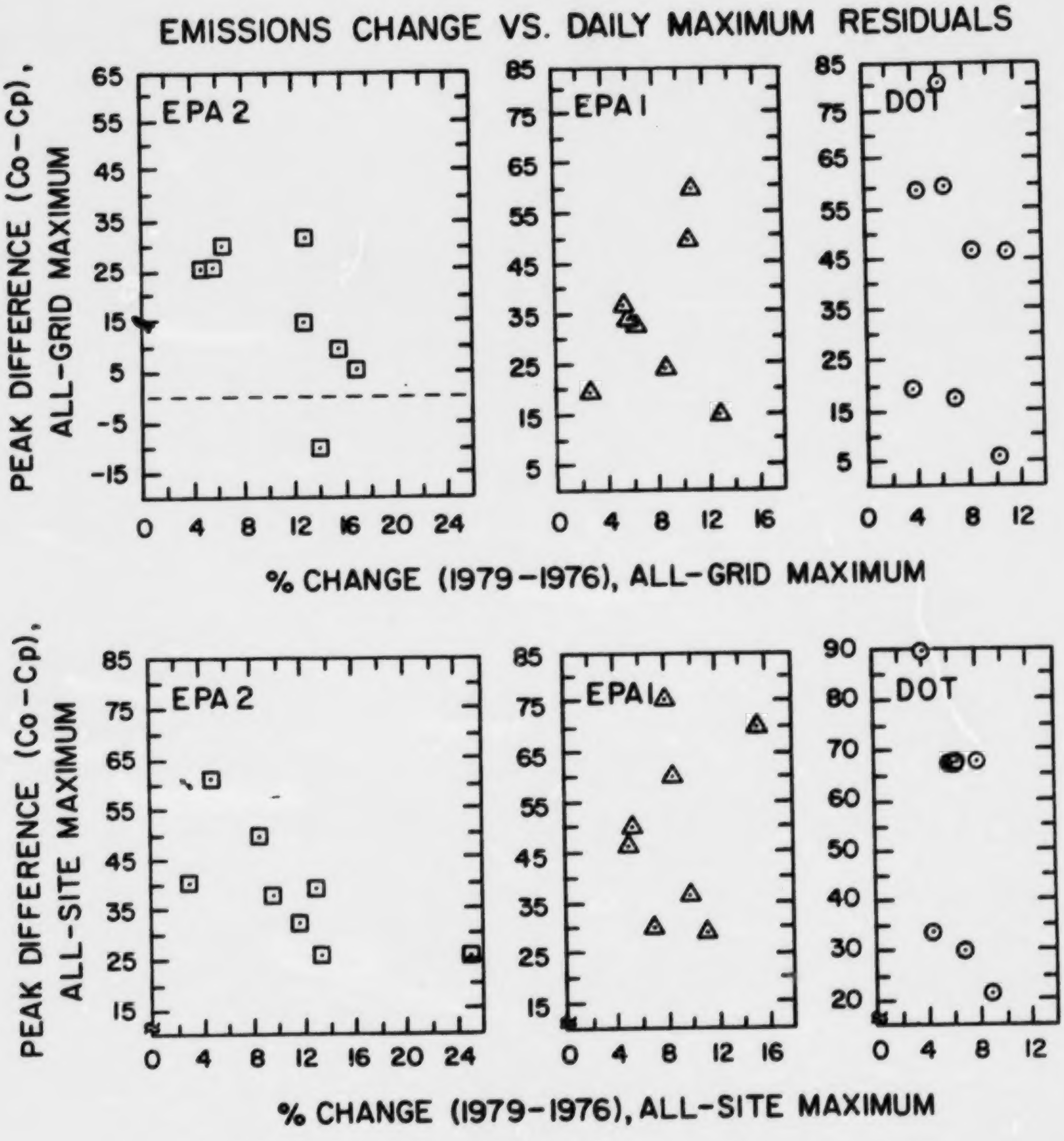


FIGURE 43: Emissions Change Results vs Daily Maximum Residuals for EPA2 and DOT Models

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA 450/4-83-021	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE Evaluation of Performance Measures for an Urban Photochemical Model		5. REPORT DATE July 1983
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Robin L. Dennis, Mary W. Downton and Robbi S. Keil		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS National Center for Atmospheric Research Environmental and Societal Impacts Group P.O. Box 3000 Boulder, Colorado 80307		10. PROGRAM ELEMENT NO. A13A2A
		11. CONTRACT/GRANT NO. AD-49-F-0-167-0
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency Office of Air Quality Planning and Standards Monitoring and Data Analysis Division (MD-14) Research Triangle Park, North Carolina 27711		13. TYPE OF REPORT AND PERIOD COVERED
		14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES Henry S. Cole, Project Officer		

16. ABSTRACT

A workshop conducted by the American Meteorological Society for EPA in September 1980 recommended a large set of statistical measures for use in the evaluation of air quality models. The present study was designed to test the recommended measures in an actual performance evaluation of an airshed model on data developed for Denver, Colorado. Three versions of the SAI Urban Airshed Model were examined. The study involved both an evaluation of the models and an evaluation of the statistical performance measures. The evaluation of the models had two parts--a base year case and an emissions trend case, the latter representing the use of the models for regulatory purposes. Resulting recommendations are intended to aid the future use of such models, the planning of future performance evaluations of airshed models, and the use of performance evaluation statistics.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air pollution Atmospheric models Photochemical reactions Smog Ozone Nitrogen Oxides Hydrocarbons	Urban Airshed Model SAI Airshed Model Carbon-Bond Mechanism Denver	

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 219
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE